Couplage entre introduction et réparation des cassures double brin pendant les réarrangements programmés du génome de Paramecium tetraurelia

Antoine Marmignon

Centre de génétique moléculaire
Bâtiment 26
1 avenue de la Terrasse
91190 Gif sur Yvette

Thèse de doctorat de l'université Paris Sud XI
Ecole doctorale : Gènes, Génomes, Cellules

Présentée par Antoine MARMIGNON

Couplage entre introduction et réparation des cassures double brin pendant les réarrangements programmés du génome chez *Paramecium tetraurelia*.

Soutenance le 27 septembre 2013
Devant le jury composé de :

Président du jury : Jean Cohen
Rapporteur : Jean Baptiste Charbonnier
Rapporteure : Gaelle Legube
Examinatrice : Ariane Gratias-Weill
Directrice de thèse : Mireille Bétermier

Remerciements

On m’avait dit que les remerciements s’écrivaient en dernier, tout à la fin, et que c’était très facile, le tout étant de ne pas provoquer de crise diplomatique en oubliant quelqu’un. Rien de tel pour nous mettre la pression, à moi et ma mémoire de poisson rouge. J’ai déjà du mal à retenir les dates d’anniversaire de ma famille alors… pour tout ceux que je vais oublier, car il y en aura probablement, je m’excuse par avance et vous remercie mille fois également.

En premier lieu, je tiens à remercier les membres de mon jury qui ont accepté de juger mon travail de thèse. Certains ont eu besoin d’être convaincu. Pour obtenir leur accord, je me suis engagé à ne pas terminer la soutenance par un paint-ball endiablé ou un jeu de rôle moyenâgeux. Je tiendrais promesse.

Je souhaite également remercier les organismes qui ont financé ma recherche, le ministère d’abord, puis la fondation ARC. L’idéal serait que la fondation ARC reste dans ces bonnes dispositions et porte un regard bienveillant sur mes futures demandes d’argent, ca faciliterait grandement la poursuite de ma carrière dans un futur très très proche.

S’il est une personne que je n’oublierai pas dans ces remerciements, c’est bien Mireille…

Mais revenons un peu dans le temps d’abord. Et intéressons nous à moi, étudiant de M2, il y a quelques années. Je dois faire un séminaire bibliographique dans le cadre de ma formation à l’université. Manque de bol (ou plutôt coup de bol) le sujet choisi dans la liste est trop plébiscité et comme je suis bonne pate, j’accepte un autre sujet, délaissé jusqu’ici. Cela

concerne les réarrangements programmés du génome et les marques épigénétiques chez les ciliés, encadré par Sandra Duharcourt (C'est ainsi que j'ai fait mes premiers pas dans le monde des ciliés. Et non ce n'est pas encore Mireille, mais merci Sandra pour m'avoir permis de mettre le pied dans la porte). Je trouve le système fascinant. Plus tard je devrai trouver un sujet de thèse, à l'occasion d'une discussion avec Jean-Luc Ferat au deuxième étage du CGM, je parle de ce que j'ai fait pendant le semestre, et il m'annonce qu'au rez de chaussée du bâtiment, Mireille Bétermier travaille sur les réarrangements programmés du génome de la paramécie (Merci Jean-Luc pour le tuyau). Je descends donc au rez de chaussée, passe la tête par la porte de la pièce 38 et dit que je cherche un stage avec une possibilité de thèse après. Mireille, que je rencontre pour la première fois, me répond que là tout de suite elle est occupée mais me propose de revenir le lendemain même heure pour en discuter.

Ce fut le début de mon aventure dans la paramécie et je ne regrette pas. Etre encadré par Mireille, c'est avoir une chef dynamique, patiente, toujours disponible, investie, rigoureuse, et profondément humaine (La liste est plus longue mais j'arrive déjà au bout de la première page…). C'est certainement grâce à elle et à la bonne ambiance qui règne dans le laboratoire, grâce à tous ses membres, que venir au labo le matin, et même la nuit, et le week-end (oui oui) pour pouponner des paramécies ne fut jamais une corvée.

Donc Mireille, Aurélie, Céline, Jacek, Emeline, Nathalie, Julien, Vinciane, et toutes les personnes qui sont passées au labo ou qui font ou ont fait partie de l'aile bisounours pendant ces quatre années, merci beaucoup.

Le monde de la paramécie est une petite communauté, riche en interactions, humaines et scientifiques, où il est agréable de travailler. Ainsi je remercie également tous les « gens de Paris » pour leur retours, leurs idées, leurs conseils et leur présence.

Puisque je parle d'environnement, intéressons nous au CGM, le Centre de Génétique Moléculaire pour ceux à qui cet acronyme ne signifie rien. Je ne peux pas faire une liste exhaustive de tous les gens que j'y ai croisés et que je tiens à remercier, mais rien n'aurait été possible sans chacune de ses parties, les dames de la laverie toujours souriantes ; le secrétariat qui gère beaucoup mieux l'administratif que moi, heureusement qu'ils sont tous là ; l'atelier et le magasin pour gérer les urgences ; l'ensemble des scientifiques qui font du bâtiment un incubateur à idées ; l'association étudiante, qui permet de tenir le coup dans l'adversité ; et la direction qui gère tout ca, si ma présence au Conseil de Laboratoire m'a appris quelquechose, c'est que diriger un institut de recherche n'a rien d'une partie de plaisir, je suis bien content de ne pas être en charge. Bref, à tous ces gens merci.

A tous mes amis hors du monde de la biologie, Charles, Matthieu, Louis, Hadrien, Jean-Yves, Jonathan, Jérôme, Sarah, Charlotte, Blandine, pour m'avoir aidé à tenir le coup (oui oui je sais j'ai dit que venir au labo n'était pas une corvée, mais je n'ai jamais dit qu'être thésard était facile), qui me changent les idées, m'emmènent voir des navets au cinéma pour déconnecter les neurones (pas que, je vais aussi voir des choses plus intellectuelles), me permettent d'avoir une vie en dehors du labo, je tiens à vous dire merci, merci beaucoup.

A ma famille, je sais que le cursus universitaire n'a pas toujours semblé à mes parents la voie idéale pour me garantir une vie longue et prospère. Mais c'est grâce à eux que j'ai développé ma curiosité de tout, mon goût pour la science et les questions existentielles qui m'ont fait m'épanouir dans la pratique de la recherche.

JE NE REMERCIE PAS, les paramécies asynchrones qui m'ont valu de nombreuses nuits de veille pour rien au laboratoire, à les couver d'un regard implorant pour qu'elles aient

la bonne volonté de démarrer les processus sexuels toutes en même temps. De toute façon, celles qui n'ont pas obéi ont été bien punies… je les ai javellisées sans remords (C'est le coté obscur de la Science).

JE NE REMERCIE PAS, ma voiture. Si l'avoir m'a permis de continuer à pratiquer l'équitation dans mon club préféré tout au long de ma thèse, la tentation de l'utiliser abusivement a été trop forte et se ressent sur mon tour de taille (On pourrait croire que c'est de la mauvaise foi et que je devrais ne m'en prendre qu'à moi ; on aurait raison).

Ca y est l'inspiration me fait défaut… Encore une fois, à tous ceux que j'oublie, ce n'est pas que vous n'avez pas compté ; blâmez ma mémoire défaillante, et recevez tout de même tous mes remerciements.

Table des matières

Liste des figures

Introduction

Résultats

Discussion

Liste des abréviations

A.

A : adénine
aa : acide aminé
ADN : acide désoxyribonucléique
alt-NHEJ : voie du non homologous en joining alternative
ARN : acide ribonucléique
ARNi : interférence ARN
ARNm : ARN messager

B.

BET : bromure d'ethidium

C.

C : cytosine
CDB : cassure double brin de l'ADN
CSB : cassure simple brin de l'ADN

D.

DAPI : 4',6'-diamino-2-phenylindole (intercalant de l'ADN)
DNA-PKcs : sous unité catalytique de la protéine kinase dépendante de l'ADN
Da : Dalton

E.

EJ : End joining, recollement des extrémités

G.

G : guanine
GFP : green fluorescent protein, protéine fluorescente verte de la medure

H.

HMG : high mobility group

I.

IES : internal eliminated sequence
IP : immunoprécipitation

J.

J : segment sous-génique de jonction dans la recombinaison V(D)J
JH : jonction de Holliday

K.

kb : kilobase
KO : Knock Out

M.

MAC : macronoyau, macronucléaire
MDS : chez Oxytricha, macronuclear destined sequences
mic : micronoyau, micronucléaire
MMEJ : microhomology mediated end joining
MRN : complexe Mre11-Rad50-Nbs1
MRX : complexe Mre11-Rad50-Xrs2
mt : type sexuel

N.

NHEJ : non homologous end joining.

O.

3’-OH : groupement hydroxyl

P.

pb : paire de bases
PCR : polymerase chain reaction
5’-PO4 : groupement phosphodiester.

R.

RH : recombinaison homologue
RSS : séquence signale de recombinaison

S.

SSA : single strand annealing
SDSA : synthesis dependent strand annealing

T.

T : thymine
TdT : terminal deoxynucléotidyl transférase

U.

U : uracile
UV : ultra violet

V.

V : segment sous génique de variabilité dans la recombinaison V(D)J

W.

WGD : duplication globale du génome
WT : sauvage

X.

XRCC : groupe de complémentation de cellules sensibles aux radiations ionisantes.

INTRODUCTION

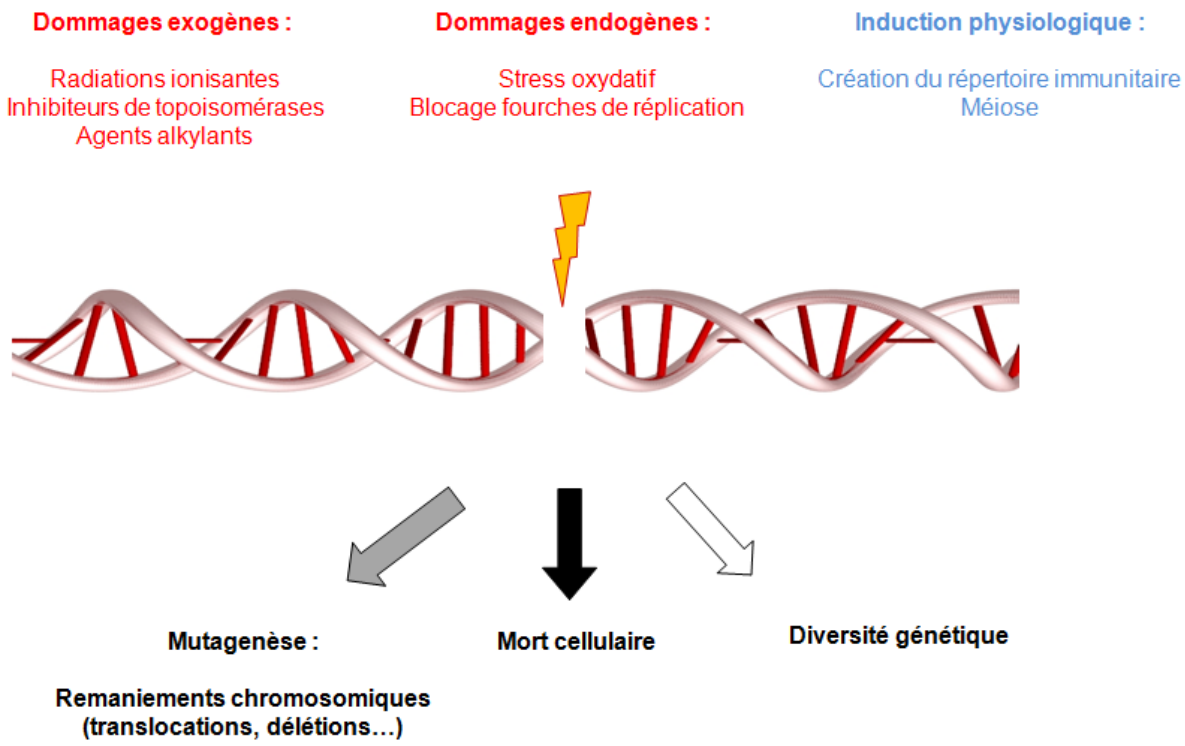


Figure 1 : Représentation schématique des différentes sources de dommages de l'ADN et des conséquences possibles. En rouge apparaissent les dégâts accidentels, en bleu les dégâts programmés.

Les cassures de l'ADN

L'ADN est le support de l'information génétique, il permet l'expression des nombreuses protéines et structures cellulaires nécessaires à la vie des organismes, qu'ils soient unicellulaires ou plus complexes et pluricellulaires. L'ADN est transmis à la descendance, il doit donc être suffisamment stable pour être conservé de génération en génération. Cependant l'ADN est une molécule complexe qui est exposée à de nombreuses sources de dommages (figure 1). Il existe de nombreux types de dommages qui peuvent être introduits de façon accidentelle ou programmée :

Les cassures accidentelles. L'ADN de par son rôle central est constamment manipulé dans la cellule, pour être transcrit, répliqué, etc… Les machineries responsables de ces mécanismes peuvent connaitre des ratés qui vont entrainer la formation de cassures de l'ADN. Les radiations ionisantes, de nombreux composés chimiques, et même les produits du métabolisme de la cellule, les espèces réactives de l'oxygène, peuvent également endommager l'ADN.

A ces cassures accidentelles s'ajoutent les cassures dites programmées. Ces cassures ne sont pas des accidents mais font partie d'un processus « volontaire » et programmé, et souvent indispensable pendant le développement ou la différenciation cellulaire. Par exemple, lors de la méiose, des cassures double brin de l'ADN sont introduites par la nucléase Spo11 de façon programmée et sont nécessaires à la ségrégation des chromosomes homologues. Un autre exemple fameux est le cas de la recombinaison V(D)J, des cassures double brin sont introduites et sont essentielles à la formation des anticorps dans les lymphocytes et à la création d'un large répertoire immunitaire. Ces cassures programmées sont donc essentielles au fonctionnement normal du système immunitaire. Ces deux exemples seront détaillés plus loin.

Pour résister à tous ces outrages, l'évolution a sélectionné de nombreux mécanismes pour réparer les multiples dommages qui peuvent affecter la molécule d'ADN. Dans ce mémoire je m'intéresserai particulièrement à l'introduction programmée de cassures double brin et leur réparation c'est pourquoi je concentrerai mon introduction sur les voies majeures de réparation des cassures double brin de l'ADN.

Réparation des cassures double brin de l'ADN

Comme leur nom l'indique, les cassures double brin de l'ADN coupent les deux brins de la double hélice d'ADN et sont donc susceptibles de séparer une molécule en deux parties. Ces cassures peuvent être le produit entre autres d'incidents de réplication, ou d'exposition à des radiations ionisantes ou des agents chimiques. Ces cassures sont extrêmement dangereuses pour la cellule car elles sont génératrices de mutations, de perte de parties de chromosomes, ou de translocations entre chromosomes. Dans une cellule humaine, une cassure double brin de l'ADN non ou mal réparée est potentiellement létale, si elle interrompt un gène essentiel par exemple.

La littérature scientifique décrit deux voies majeures de réparation des cassures double brin, la Recombinaison Homologue (RH) et la voie du Non Homologous End Joining (NHEJ). On découvrira qu'il existe d'autres voies apparentées qui peuvent être mobilisées selon les circonstances.

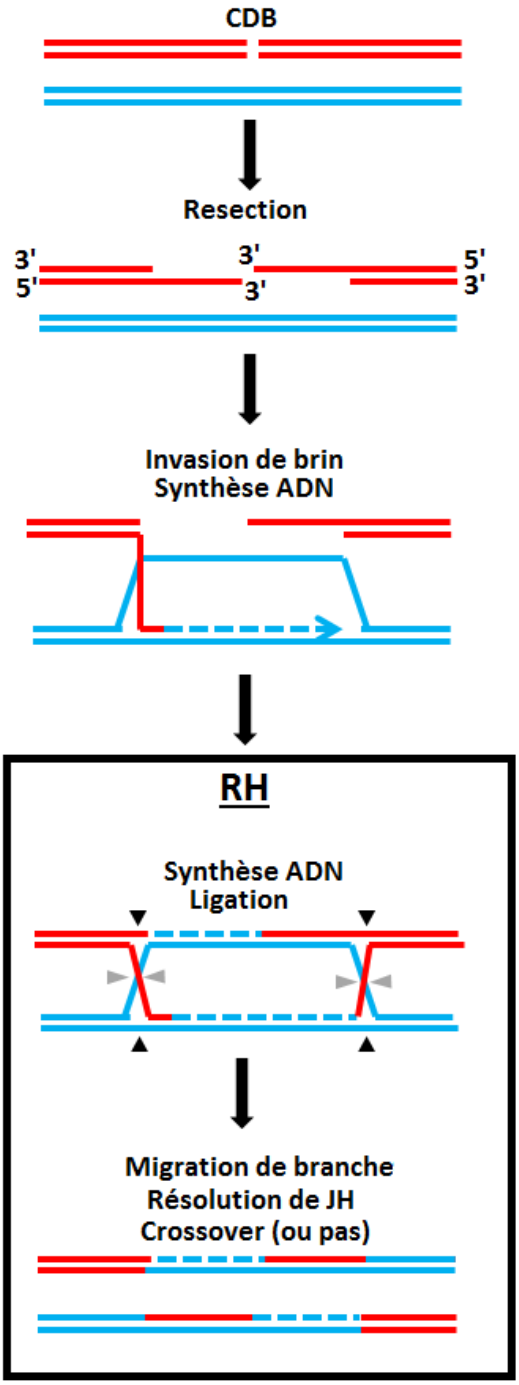


Figure 2 : Représentation schématique de la voie de recombinaison homologue et de ses différentes étapes. L'ADN cassé est en rouge, l'ADN homologue est en bleu. Les flèches en pointillés représentent les étapes de synthèse d'ADN. Les triangles en noir et gris représentent les différentes façon de résoudre les jonctions de Holliday.

La recombinaison homologue :

La recombinaison homologue est une voie, probablement apparue très tôt au cours de l'évolution, très conservée de la bactérie jusqu'à l'homme. Elle permet la réparation précise des cassures double brin en utilisant un ADN homologue qui sert de matrice de réparation, il sera recopié pour remplacer l'ADN au niveau de la cassure (Li and Heyer, 2008). Lorsqu'une cassure double brin est introduite et détectée, l'extrémité est prise en charge par un complexe de résection. Il va préparer les extrémités en dégradant l'extrémité 5' de l'ADN pour libérer un ADN simple brin en 3' de plusieurs centaines de bases de long. L'ADN simple brin va envahir l'ADN complémentaire et va servir d'amorce pour la réplication de l'ADN. La ligation des brins néosynthétisés avec les ADN matrices entraine la formation de jonctions de Holliday (JH) qui seront résolues, entrainant éventuellement l'échange de séquences alléliques entre les deux molécules d'ADN, on parle alors de *crossing over* (figure 2 et 3).

Les principaux acteurs de la recombinaison homologue

Etapes de la recombinaison homologue	**Principales protéines procaryotes**	**Principales protéines eucaryotes**
Coupure double brin de l'ADN	Aucune	Spo11 pendant la méiose
Génération d'une extrémité simple brin en 3'	RecBCD	Mre11/Rad50/Xrs2 (Nbs1) Sae2 (CtIP), Exo1
Invasion de l'ADN complémentaire	RecA	Rad51 Dmc1 pendant la méiose
Jonctions de Holliday et migration de branche	RuvAB	Mus81-Eme1
Résolution des jonctions de Holliday	RuvC	Slx1-Slx4 BLM, GEN1

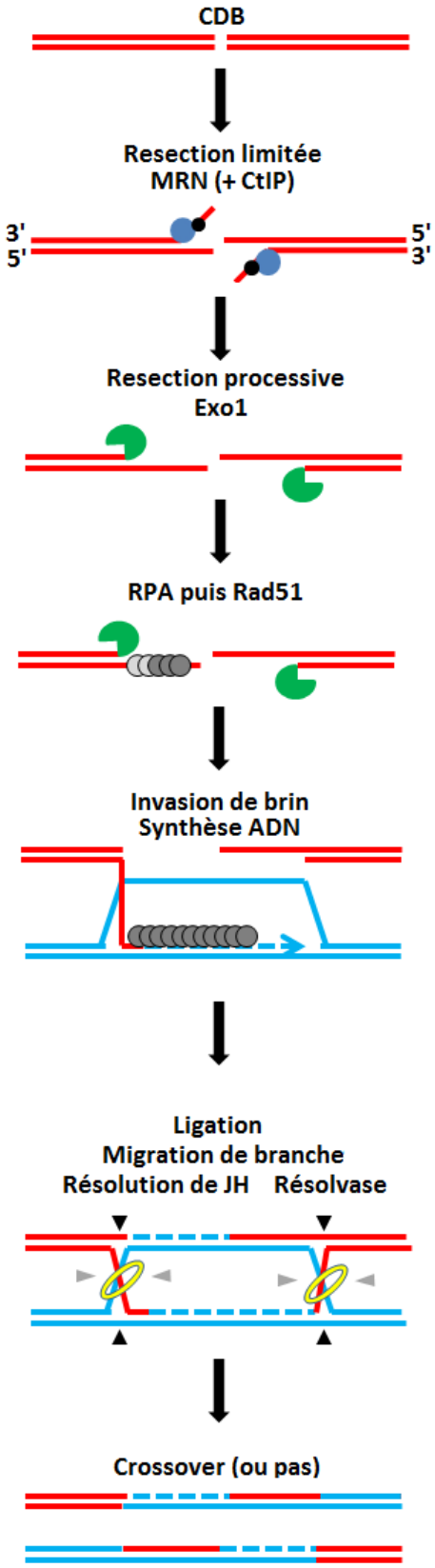


Figure 3 : Représentation des principaux acteurs protéiques de la recombinaison homologue. De nombreuses autres protéines sont mobilisées mais n'aparaissent pas sur ce schéma.

Les cassures de l'ADN qui vont être réparées par recombinaison homologue peuvent être accidentelles ou programmées. **La protéine Spo11** est spécifique de la méiose. Il s'agit d'une endonucléase, agissant sous forme de dimère, indispensable pour l'introduction des cassures double brin programmées pendant la méiose. (Baudat and de Massy, 2004; Longhese et al., 2009) Après avoir coupé l'ADN, la protéine est liée à l'ADN, formant un obstacle à la résection, et doit être retiré pour réparer par recombinaison homologue. C'est le rôle de CtIP

CtIP est l'homologue de la protéine Sae2 qui avait été précédemment découverte chez la levure. Cette protéine, dont les fonctions ne sont pas encore claires semble impliquée dans les étapes initiales de résection des cassures double brin. On pense que CtIP a une activité endonucléase qui lui permet d'éliminer les adduits Spo11-ADN d'une part, et d'initier la résection en association avec le complexe MRN d'autre part. Elle semble également avoir un rôle dans l'activation de point de contrôle du cycle cellulaire, qui permettront de prendre le temps de réparer la cassure avant la reprise normale du cycle cellulaire (Huertas and Jackson, 2009; Mimitou and Symington, 2009).

Les protéines Mre11, Rad50 et NBS1 (ou Xrs2) forment le complexe MRN chez les eucaryotes supérieurs (MRX chez la levure). Les KO de ces gènes ont montré qu'ils étaient essentiels pour la croissance et la viabilité. Ces protéines font partie des premiers acteurs à se localiser au niveau des cassures double brin et facilitent l'entrée dans la voie de recombinaison homologue mais semblent impliquées également dans la voie NHEJ. *In vitro,* Mre11 a une activité endonucléase 5' vers 3' et exonucléase de 3' vers 5'. *In vivo* cette protéine coopère avec CtIP pour initier la résection des extrémités ADN. Rad50 fait partie de la famille des protéines SMC : elle est impliquée dans la maintenance de la structure des chromosomes. Nbs1 ou son homologue Xrs2 de levure sont des protéines qui interagissent

avec des régulateurs du cycle cellulaire et des protéines de checkpoint, comme par exemple la kinase ATM. (Deriano et al., 2009; Lisby et al., 2004)

Exo1 est une exonucléase 5' vers 3' impliquée dans la maturation des extrémités d'ADN pour la recombinaison homologue. Elle permet d'étendre la résection sur de larges distances, d'autres protéines, nucléases et hélicases, comme CtIP et Mre11 ayant initié le processus (Mimitou and Symington, 2008, 2009).

Rad51 va permettre l'invasion de l'ADN double brin homologue**.** La protéine RPA, homologue de SSB chez les bactéries va se fixer sur l'ADN simple brin sortant en 3' généré par la résection du brin 5'. Rad51 va la remplacer et recouvrir l'extrémité sortante 3' qui va envahir la molécule ADN homologue intègre. De nombreuses autres protéines sont impliquées dans cette étape d'invasion, notamment les paralogues de Rad51, Rad51B, Rad51C, Rad51D et la protéine XRCC2. Leur rôle n'est pas encore tout à fait clair, ils semblent promouvoir la recombinaison homologue en favorisant la formation du filament Rad51 sur l'ADN recouvert par RPA. Spécifiquement pendant la méiose, la protéine Dmc1 est un autre partenaire de Rad51. (Li and Heyer, 2008)

Après l'action de polymérases, nucléases, la ligation puis la résolution des jonctions de Holliday a lieu, avec ou sans *crossing over*. (Li and Heyer, 2008)

Mécanisme programmé utilisant la recombinaison homologue : la méiose

Des cassures double brin sont introduites par l'endonucléase Spo11 pendant la méiose et sont réparées par recombinaison homologue. L'absence de Spo11 abolit l'introduction de cassures

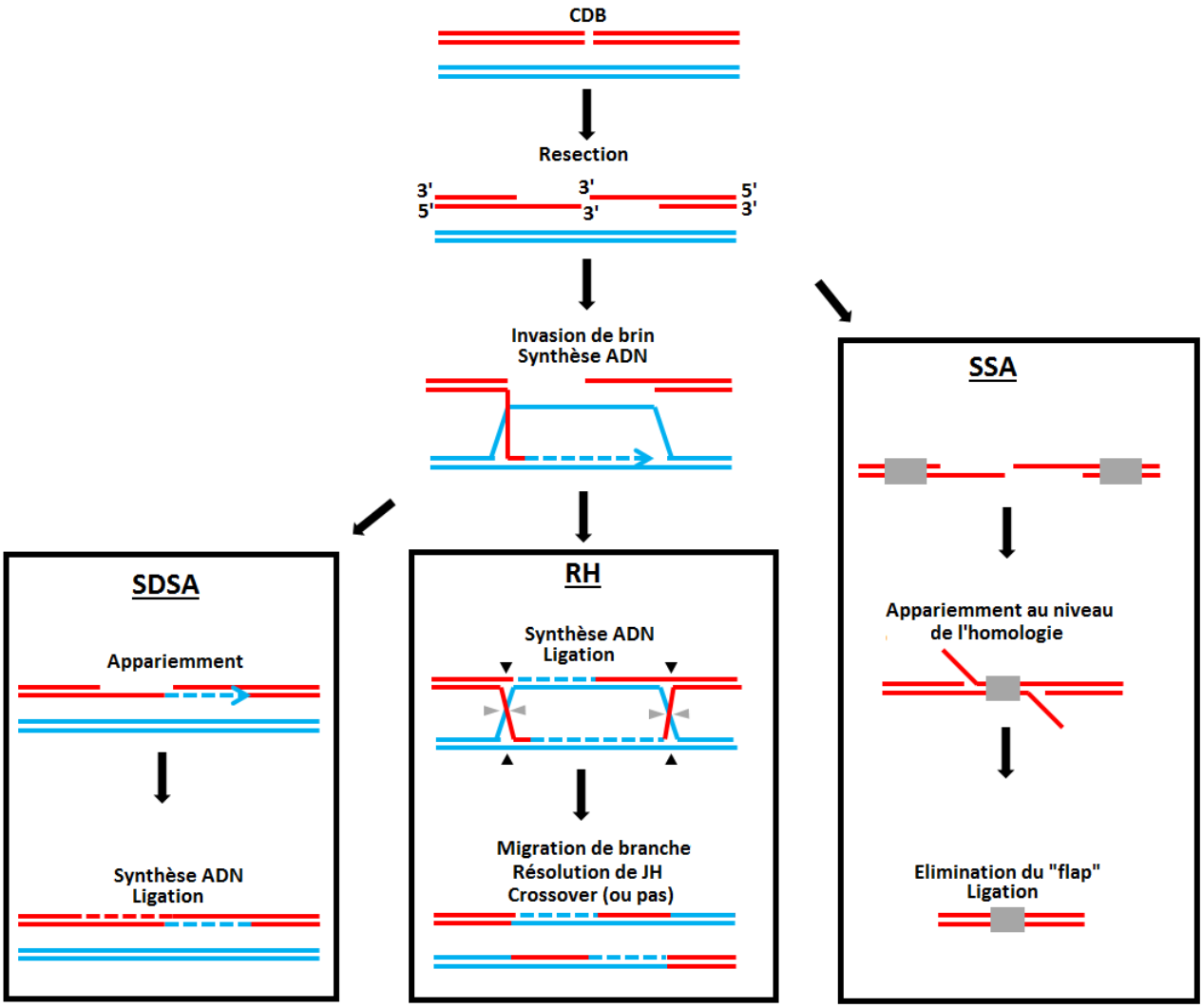


Figure 4 : Représentation schématique de la voie de recombinaison homologue et des voies apparentées. L'ADN cassé apparait en rouge tandis que l'ADN homologue est en bleu. Les rectangles gris indiquent une zone d'homologie.

programmées pendant la méiose mais cela est létal. Ce résultat était un peu paradoxal dans le sens où on imagine qu'il est dans l'intérêt de la cellule de limiter au maximum l'introduction de cassures double brin dans son génome. Cependant des études ont montré que les cassures introduites pendant la méiose permettent l'appariement des chromosomes homologues et la recombinaison avant leur ségrégation lors de la première division de méiose. Lorsque les cassures ne se font pas et que la recombinaison est empêchée, les chromosomes sont mal alignés et cela entraine des problèmes de ségrégation. Les cellules filles se retrouvent avec des chromosomes excédentaires ou manquants, entrainant des problèmes associés à la polyploïdie ou l'aneuploïdie. (Baudat and de Massy, 2004)

Les voies apparentées à la recombinaison homologue

Synthesis Dependent Strand Annealing (SDSA)

Le SDSA est une voie proche de la recombinaison homologue, les premières étapes sont exactement les mêmes avec résection du brin 5', invasion de l'ADN homologue puis synthèse d'ADN jusqu'à atteindre une zone de microhomologie. Au niveau de l'extrémité 3' de l'autre coté de la cassure, le brin néosynthétisé va alors s'apparier à la séquence homologue correspondante et poursuivre la synthèse d'ADN pour réparer la cassure sans générer de jonctions de Holliday (figure 4). Cependant la réparation peut entrainer des délétions.

Single Strand Annealing (SSA)

Le SSA est une voie apparentée à la recombinaison homologue dans le sens ou elle débute de la même façon par la résection du brin 5', à l'image du SDSA, mais c'est la seule étape qu'elle partage avec la recombinaison homologue : il n'y a pas d'invasion de brin sur une

autre molécule d'ADN duplexe, c'est pourquoi la protéine Rad51 n'est pas requise. La résection se poursuit à longue distance jusqu'à dégager des séquences d'homologie pouvant aller jusqu'à plusieurs centaines de paires de bases. Ces séquences homologues sont appariées grâce à la protéine Rad52. Les extrémités 3' non appariées générées sont éliminées et la ligation termine la réparation de la cassure (figure 4). Ce mécanisme entraine donc l'introduction de délétions qui peuvent s'étendre sur de longues distances.

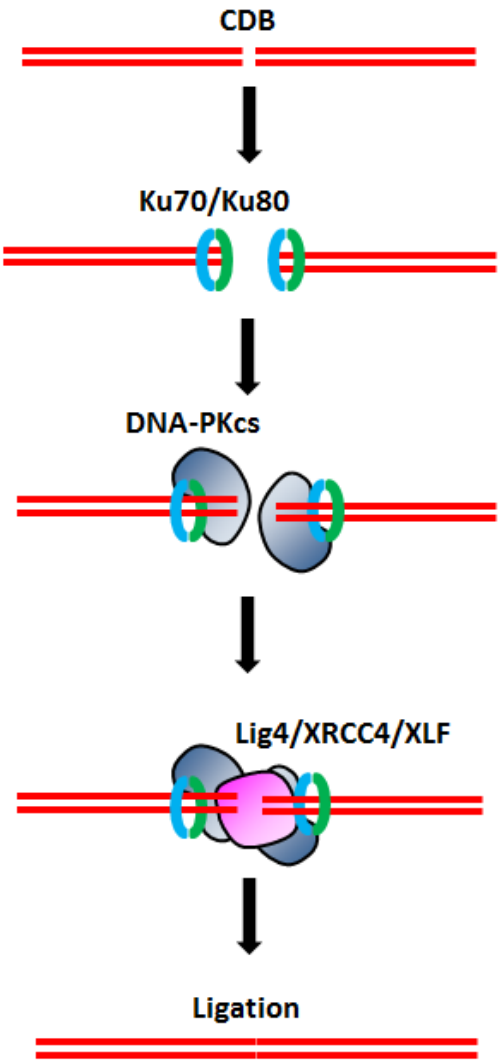


Figure 5 : Les différentes étapes et les principaux acteurs de la voie de Non Homologous End Joining. Les protéines accessoires et les facteurs de maturation éventuels n'apparaissent pas.

La voie de Non Homologous End Joining :

Moins conservée que la RH, la voie de Non Homologous End Joining (figure 5) est présente chez certaines bactéries et tous les organismes eucaryotes. Comme son nom l'indique, cette voie diffère de la RH par le fait que les extrémités d'ADN sont directement religuées sans nécessité d'une matrice ADN homologue (Lieber et al., 2008). C'est évidemment la seule voie possible lorsqu'aucun ADN homologue n'est disponible dans le noyau (ou la cellule pour les bactéries), comme cela est le cas dans les cellules haploïdes, à l'exemple des spores de levures. Si les extrémités d'ADN sont parfaitement compatibles, la réparation sera extrêmement précise. La génération d'extrémités compatibles peut nécessiter des étapes de maturation. Dans ce cas, il arrive souvent que la jonction de réparation soit refermée de façon imprécise, pouvant entrainer la perte de nucléotides ou l'insertion de séquences supplémentaires au point de jonction. De plus une mauvaise réparation dans le cas d'une voie NHEJ défaillante peut mener à des translocations plus nombreuses ou favoriser les fusions de télomères (Guirouilh-Barbat et al., 2004). Cela est souvent observé dans les cellules cancéreuses.

Les principaux acteurs du NHEJ

Complexe du NHEJ	**Protéines procaryotes**	**Protéines eucaryotes**
DNA-PK	YkoV (Ku)	Ku70/80 DNA-PKcs
Maturation		Artemis TdT
Ligation	YkoU = LigD	LigIV-XRCC4-XLF

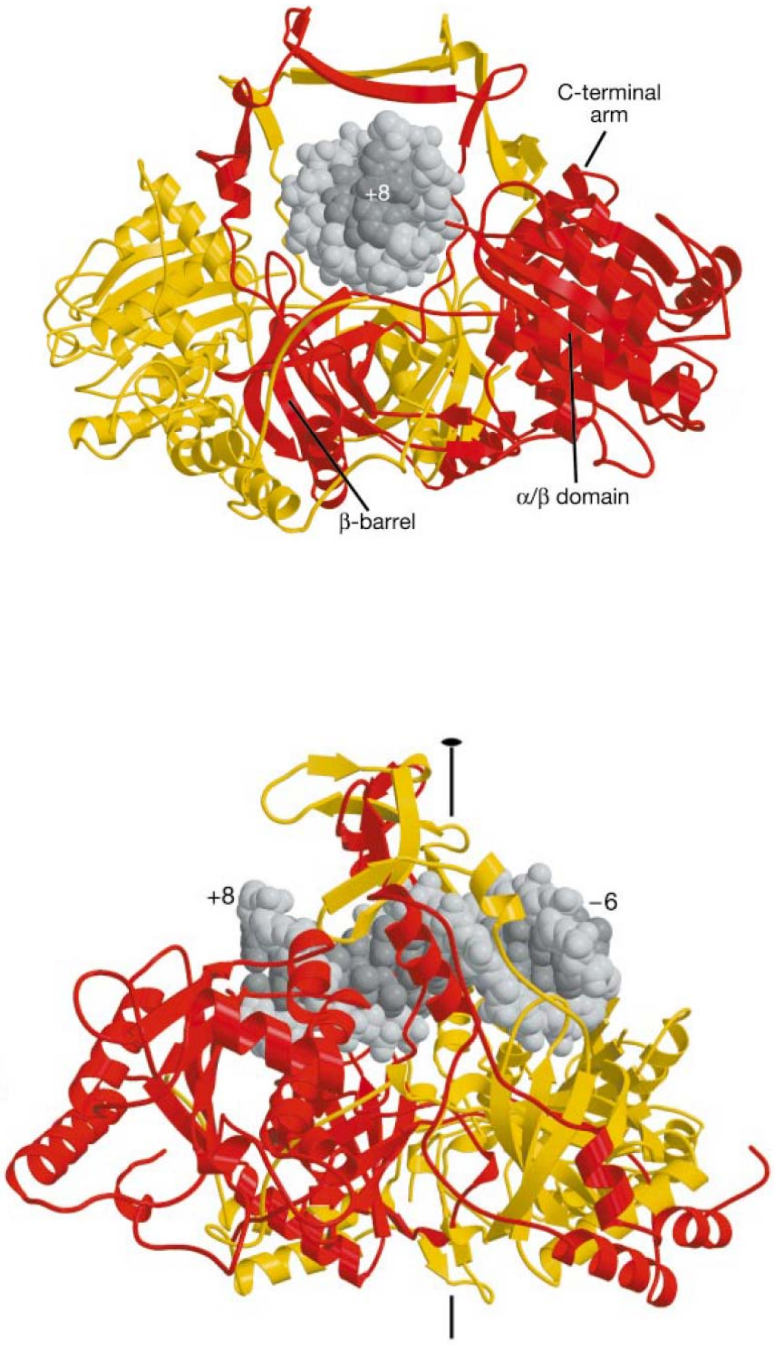


Figure 6 : Structure de l'hétérodimère Ku70/Ku86 lié à l'ADN sous deux angles différents. On observe facilement la structure en anneau qui encadre l'ADN La partie Cter de Ku80 n'apparait pas. Ku70 est en jaune, Ku80 en rouge et l'ADN en gris (Walker et al., 2001)

L'hétérodimère Ku70/80 est un des acteurs majeurs du NHEJ, il a été décrit comme une plate forme pour le recrutement des acteurs de la voie (Weterings and Chen, 2008). L'hétérodimère Ku a été initialement identifié comme un autoantigène présent dans le sérum de patients souffrants de maladies autoimmunes (MIMORI et al., 1981). La purification de cet antigène a montré qu'il s'agissait d'un complexe de deux protéines de 86 et 70 kDa (Ku86 et Ku70, respectivement). L'hétérodimère a une très grande affinité pour les extrémités d'ADN double brin, sans spécificité de séquence. Il est nécessaire pour le recrutement de DNA-PKcs au niveau des cassures. La résolution de la structure 3D de la protéine montre que l'hétérodimère forme un anneau qui peut entourer l'ADN et probablement le protéger de la dégradation par des nucléases. La protéine est composée de deux sous unités relativement symétriques, elles mêmes divisées en trois grands domaines (figure 6) (Walker et al., 2001).

L'association des domaines centraux β barrel des deux sous unités forme le berceau qui va encadrer l'ADN.

Les domaines αβ ne sont pas ou très peu impliqués dans l'interface entre les deux sous-unités ou dans l'interaction avec l'ADN, mais il est supposé que ces domaines sont importants pour les interactions avec d'autres protéines de réparation.

Les domaines Cter des deux sous unités diffèrent beaucoup. Celui de la protéine Ku80 forme un long bras moins structuré que le cœur, qui pourrait interagir avec d'autres protéines. Par exemple, l'extrémité du Cter de Ku80 interagit avec la protéine DNA-PKcs et la présence d'un motif spécifique au niveau des acides aminés terminaux est une signature de la présence de DNA-PKcs dans l'organisme considéré (Singleton et al., 1999). La délétion du Cter de Ku80 n'abolit pas le recrutement de DNA-PKcs au niveau des extrémités cassées, mais l'activité kinase de DNA-PKcs est réduite *in vivo*. Le Cter de Ku70 est plus court et pourrait être impliqué dans les interactions avec l'ADN. De nombreuses analyses des interactions de

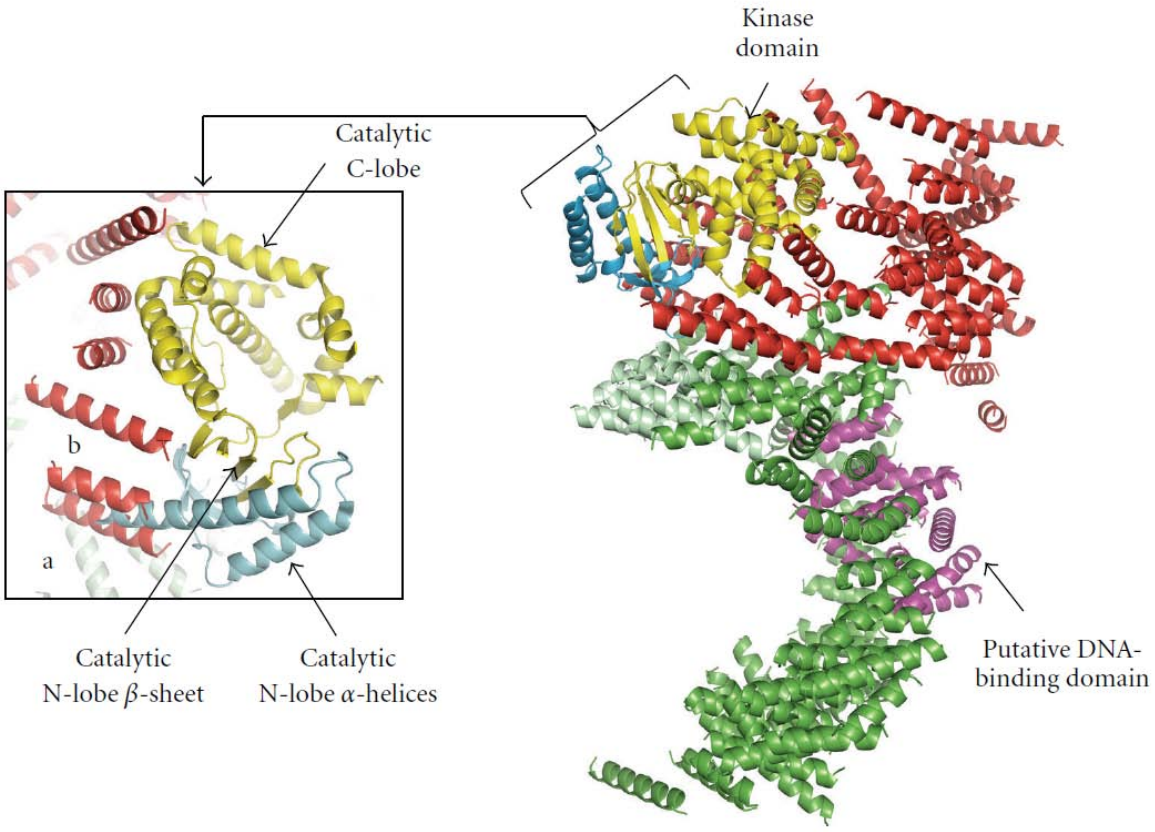


Figure 7 : Structure de la DNA-PKcs et de son domaine kinase. Modélisation du domaine catalytique de la DNA-PKcs. Les différentes couleurs représentent les différents domaines. (Ochi et al., 2010).

Ku ont été menées *in vitro* et ont montré que l'hétérodimère est capable d'interagir avec le complexe de ligation XRCC4/LigIV/XLF. Récemment, une analyse in vitro a montré que Ku avait une activité 5'-dRP/AP lyase, qui peut exciser les sites abasiques avec une forte efficacité lorsqu'ils se trouvent au niveau d'une extrémité sortante 5' d'une cassure double brin. Cela suggère que Ku pourrait être impliqué dans la maturation des cassures et important pour la fidélité de la voie NHEJ (Strande et al., 2012).

Il a été observé que Ku était impliqué dans l'addition et la maintenance des télomères au niveau des extrémités de chromosomes. Cet aspect sera abordé plus loin(Slijepcevic and Al-Wahiby, 2005).

DNA-PKcs, la sous unité catalytique de la protéine kinase dépendante de l'ADN, est une grosse protéine de 4128aa qui interagit avec Ku et l'ADN pour former le complexe DNA-PK. Comme son nom l'indique, cette protéine a une activité kinase (figure 7) et possède de très nombreuses cibles, dont elle-même. Son autophosphorylation semble jouer un rôle dans le changement de conformation du complexe DNA-PK, qui pourrait permettre l'accès de protéines de maturation des extrémités au niveau des cassures double brin ((Budman et al., 2007; Goodarzi et al., 2006). Le rôle précis de la DNA-PKcs est encore peu clair, elle semble avoir de multiples rôles dans la cellule. Elle est nécessaire pour la recombinaison V(D)J. Un KO de cette protéine chez la souris produit des souris immunodéficientes.

Le complexe de Ligation LigIV/XRCC4/XLF est responsable de l'étape finale de la réparation par la voie NHEJ. Il va permettre la ligation des deux extrémités d'ADN au niveau de la cassure. Il est formé de trois protéines dont seule la LigIV possède une activité catalytique (Ahnesorg et al., 2006; Barnes et al., 1998; Grawunder et al., 1997). Chacun des membres du complexe est indispensable pour assurer l'étape de ligation et des KO des

gènes codant ces protéines chez la souris sont létaux au stade embryonnaire. Les protéines XRCC4 et XLF interagissent et peuvent former des filaments qui pourraient favoriser le recrutement du complexe de ligation au niveau de la cassure (Ropars et al., 2011).

Les facteurs de maturation de la voie NHEJ sont encore méconnus. La protéine Artémis est impliquée dans la recombinaison V(D)J. Il s'agit d' un type particulier de réparation par NHEJ, est requise pour la résolution des structure en épingle à cheveux générées pendant l'introduction des cassures double brin programmées par les protéines RAG1 et RAG2. (Goodarzi et al., 2006; Ma et al., 2005)

Certaines polymérases de la famille X, les pol λ et pol μ peuvent remplir les *gaps* pendant le NHEJ. Par exemple, il a été montré chez la levure que la protéine Pol3, est requise pour le comblement de brèches au niveau d'extrémité 3' sortantes réparées par NHEJ lorsque la Pol4 est absente. Lorsque ces deux protéines sont absentes, il y a une forte réduction de l'efficacité de réparation par NHEJ (Chan et al., 2008).

La TdT est également une polymérase de la famille X, mais a la particularité de ne pas nécessiter de matrice. Elle permet l'addition de nucléotides aux extrémités 3' de l'ADN. Cette capacité est utilisée dans le test TUNEL qui permet d'ajouter des nucléotides couplés à des fluorophores détectables par microscopie. C'est devenu une technique de routine pour la détection de l'apoptose pendant laquelle l'ADN des noyaux est massivement cassé et dégradé. *In vivo*, la TdT est impliquée dans la recombinaison V(D)J, elle permet l'introduction de variabilité dans ce processus pour augmenter encore la diversité du répertoire immunitaire (Lu et al., 2008).

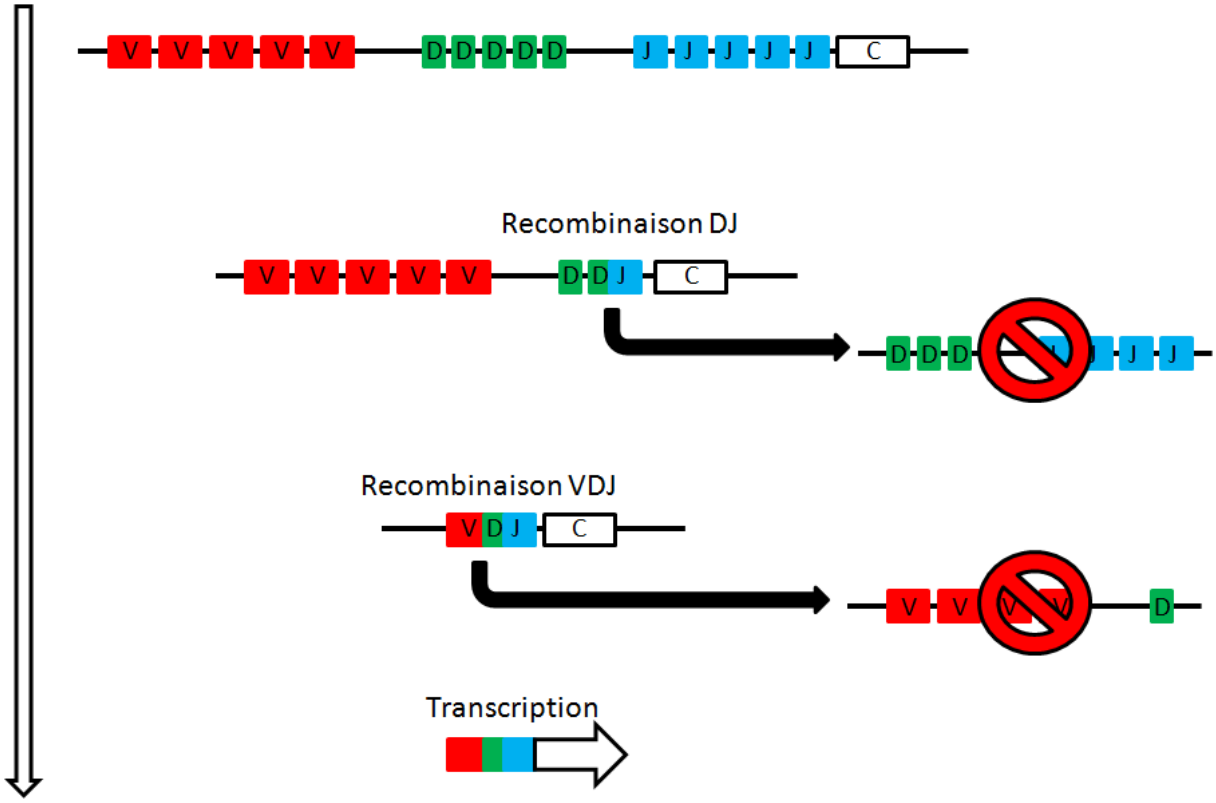


Figure 8 : Représentation simplifiée du principe de la recombinaison V(D)J permettant la diversité du répertoire immunitaire.

Mécanisme programmé utilisant la voie du NHEJ : la recombinaison V(D)J

Le séquençage du génome humain a permis de réaliser que le nombre de gènes codant pour une protéine pouvait être estimé à environ 30 000. Cependant notre système immunitaire est capables de reconnaitre virtuellement une infinité d'antigènes. Chaque anticorps reconnaissant un antigène spécifique, comment 30 000 gènes peuvent-ils coder pour une infinité d'anticorps capable de reconnaitre une infinité de motifs différents? Ce paradoxe est résolu par le mécanisme de recombinaison V(D)J. Un locus du génome est spécifiquement réarrangé pour produire des gènes variables. L'ADN de la cellule est « volontairement » cassé par les nucléases RAG1/RAG2, cela permet le réassortiment aléatoire de fragments de gènes qui permettront l'assemblage d'un gène codant pour un anticorps différent de celui de la cellule voisine (figure 8).

On peut diviser ce processus en deux grandes phases : le clivage de l'ADN par les transposases domestiquées RAG1 et RAG2 ; puis la réparation de l'ADN réarrangé par la voie NHEJ

Le complexe RAG est composé de deux protéines RAG1 et RAG2. Une protéine accessoire de la famille HMGB peut être également mobilisée. Ce complexe va se lier à une séquence 12 RSS et une 23 RSS en accord avec la règle 12/23 et former une synapse. Les RSS sont composées d'un heptamère et d'un nonamère, tout deux très conservés et séparés par une séquence intervalle de 12 ou 23 paires de bases. La formation de la synapse active avec RAG1 va introduire une cassure double brin qui libère une extrémité phosphodiester 5' et une extrémité 3' hydroxyl. Le 3'OH va attaquer l'extrémité 5'PO4 du brin antiparallèle pour former une extrémité en épingle à cheveux. Ainsi le clivage aboutit à la formation de deux épingles à cheveux du côté des séquences codantes et deux extrémités franches du coté des séquences signales RSS, non codantes.

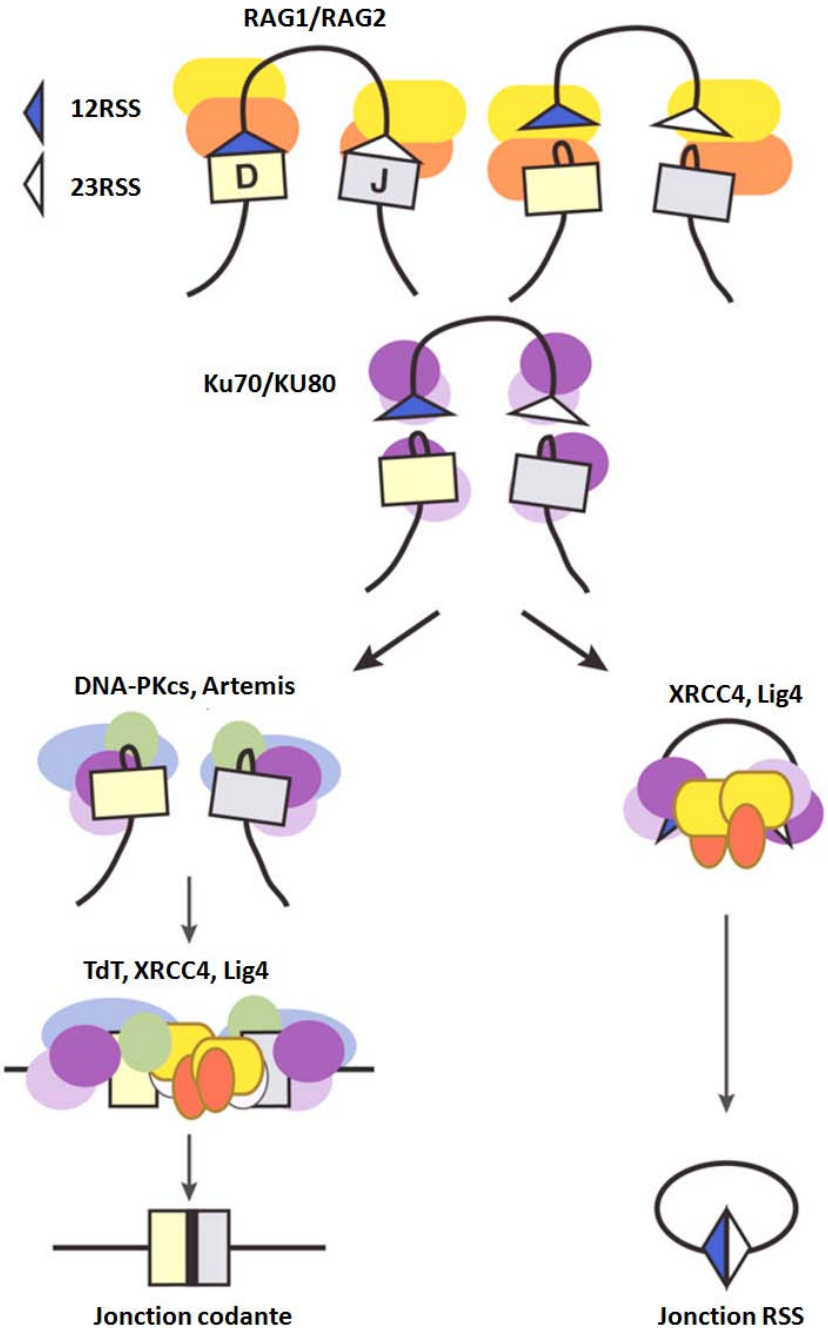


Figure 9 : Mécanismes moléculaires de la recombinaison V(D)J mis en jeu par RAG1/RAG2 au niveau des RSS. Les différentes protéines de réparation impliquées dans la réparation des jonctions codante et RSS sont indiquées également. (Bassing et al., 2002)

Pour la deuxième étape, les destins de séquences RSS et des séquences codantes sont légèrement différents. En ce qui concerne la formation de la jonction signal (ou jonction RSS), elle nécessite la présence des protéines cœurs de la voie NHEJ, Ku70/80, XRCC4 et LigIV qui vont religuer les extrémités franches. Un produit circulaire est formé et excisé sans maturation des extrémités. A l'inverse la formation de la jonction codante est plus complexe du fait de la présence des épingles à cheveux et implique le recrutement de facteurs de maturation des extrémités. L'hétérodimère Ku70/80 recrute la sous unité DNA-PKcs qui va elle-même recruter et phosphoryler Artémis. Cette protéine est une endonucléase capable d'ouvrir les structures en épingles à cheveux qui ne peuvent pas être prises en charge par le NHEJ en l'état. Une fois ces structures ouvertes, de la diversité supplémentaire va être introduite au niveau des jonctions codantes par l'action de la protéine TdT, spécifique des lymphocytes, mais aussi par l'action des polymérases μ et λ. Enfin la ligation est assurée par le complexe XRCC4-LigIV-XLF (figure 9) (Barnes et al., 1998)

Il est intéressant de noter que l'étude la voie NHEJ s'est beaucoup appuyé sur l'analyse de la recombinaison V(D)J. Cela a amené à décrire cette voie NHEJ comme sujette aux erreurs et peu précise. Cependant on observe ici que l'introduction de diversité à la jonction se fait via l'action de protéines spécialisées, n'appartenant pas au cœur de la voie NHEJ, dont le rôle est justement d'ajouter ou de retirer des nucléotides de façon non spécifique pour créer des anticorps variés permettant la reconnaissance d'une infinité d'antigènes.

Les voies apparentées au NHEJ

Les analyses des dernières années dans le domaine des voies de recollement des extrémités cassées ont permis de découvrir une voie alternative au NHEJ que l'on peut trouver dans la littérature sous de nombreux noms (MMEJ, aNHEJ, alt-NHEJ…) (Ma et al., 2003; Yu and Gabriel, 2003). Comme le nom MMEJ, pour Microhomolohy Mediated End Joining

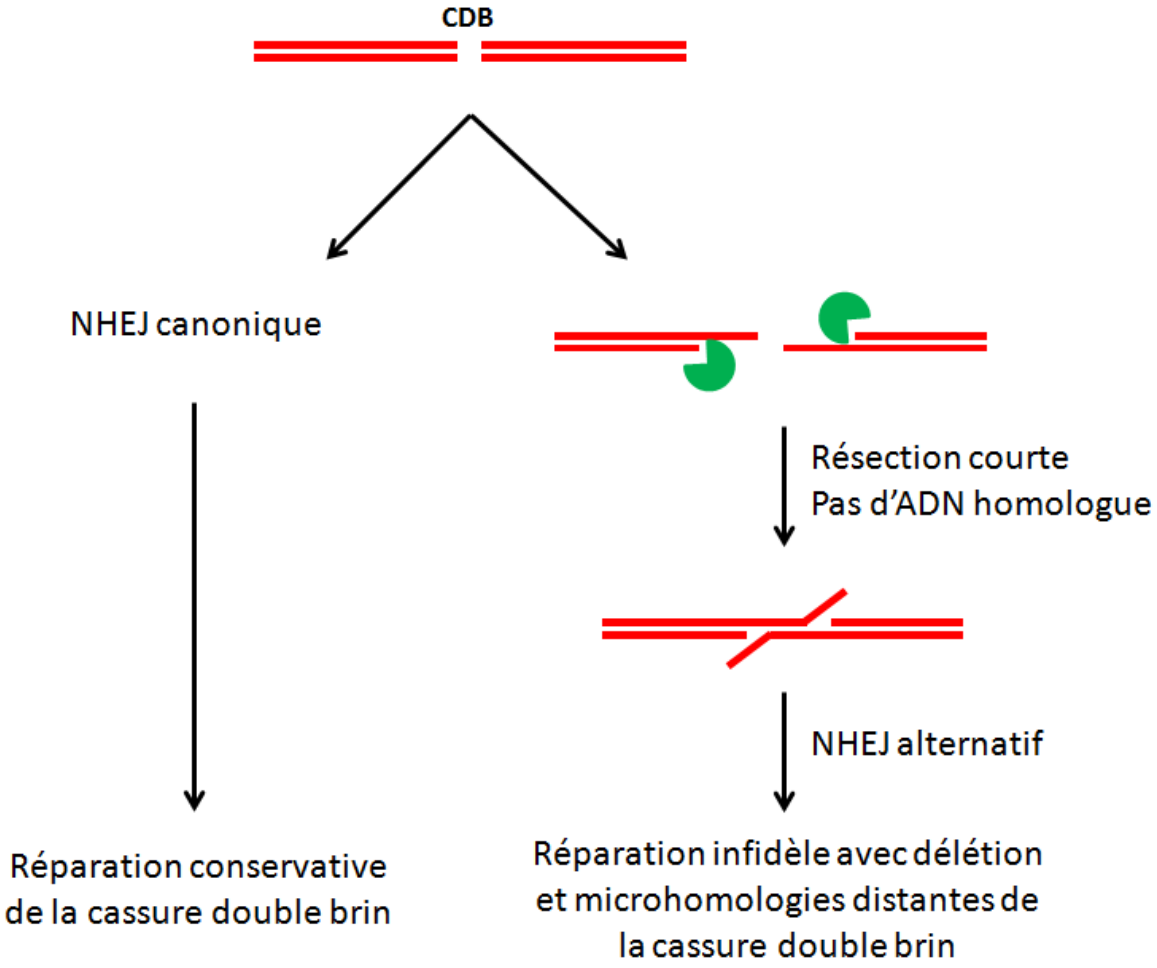


Figure 10 : Représentation schématique et simplifiée de la voie du NHEJ alternative et les conséquences au niveau de la jonction de réparation de la cassure double brin.

l'indique, cette voie est caractérisée par l'usage fréquent de microhomologies au niveau de la jonction de réparation. Elle est utilisée lorsque les protéines cœurs de la voie NHEJ sont absentes, notamment l'hétérodimère Ku70/Ku80 et la Ligase IV. Cette voie est rarement et difficilement observable en conditions normales et elle peut être vue comme une voie de secours lorsqu'un membre de la voie NHEJ est absent ou déficient (Boboila et al., 2010; Corneo et al., 2007; Wang et al., 2003). Certains indices, comme le séquençage de nombreuses jonctions de réparation, indiquent que cette voie est présente même lorsque la voie NHEJ est fonctionnelle mais la plupart du temps il est difficile de confirmer qu'il s'agit d'une voie à part entière et pas d'accident de réparation apparaissant naturellement.

Les détails de cette voie en termes d'acteurs protéiques et d'intermédiaires moléculaires sont encore largement méconnus. Cette réparation nécessite des microhomologies et nécessitent donc la résection pour les découvrir (figure 10). Le complexe MRN pourrait être impliqué dans la maturation des extrémités entrainant la perte de séquence au niveau de la jonction de réparation. Des zones de microhomologie entre 2 et 10 nucléotides seraient nécessaires pour l'appariement des brins et l'étape finale de ligation assurée par la Ligase III. (Wang et al., 2005; Wang et al., 2006)

Quelle voie choisir pour réparer une cassure double brin de l'ADN?

Une question capitale en ce qui concerne la réparation des cassures double brin est de savoir ce qui détermine la voie de réparation qui va être utilisée par la cellule. NHEJ canonique pouvant être précis, NHEJ alternatif caractérisé par des délétions, ou recombinaison homologue, généralement fidèle mais nécessitant une matrice homologue.

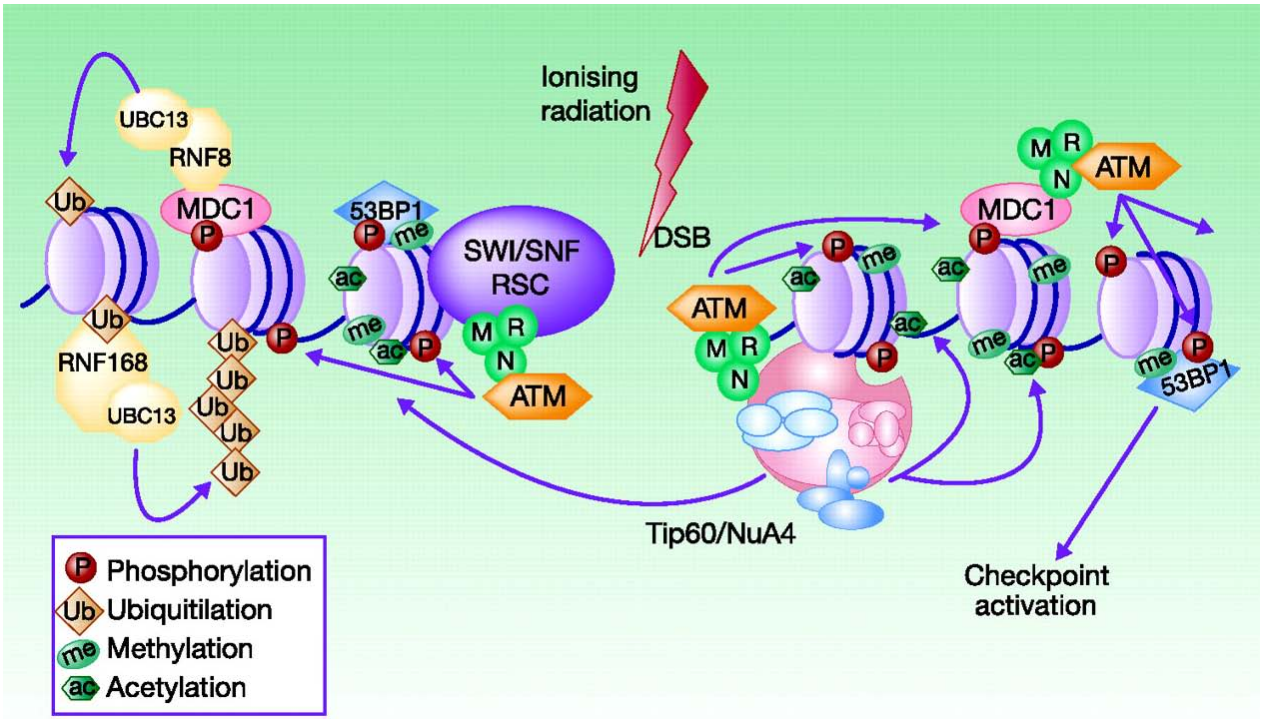


Figure 11 : Modifications épigénétiques en réponse aux cassures double brin de l'ADN. (Rossetto et al., 2010)

De façon évidente, les déterminants les plus importants est le stade du cycle cellulaire. Si l'on considère une cellule eucaryote, le NHEJ est actif à tous les stades du cycle cellulaire tandis que la recombinaison homologue n'est active que durant la réplication de l'ADN lorsqu'une chromatide sœur non cassée est disponible comme matrice (Saintigny et al., 2007; Wohlbold and Fisher, 2009). Si à l'inverse on considère une cellule de levure comme *S. cerevisiae*, celle-ci privilégie la recombinaison homologue et n'utilise la voie NHEJ que lors de sa phase haploïde. (Daley et al., 2005; Wohlbold and Fisher, 2009)

Un autre facteur important est la résection ou non des extrémités ADN. La plupart des modèles actuels indiquent que les extrémités ciblées par l'hétérodimère Ku70/80 ne sont pas ou peu maturées et orientées vers la voie NHEJ tandis que les résections plus larges initiées par la protéine CtIP, les nucléases Mre11, Exo1 et d'autres nucléases vont favoriser le chargement de Rad51 sur l'ADN simple brin et donc orienter vers une réparation par la recombinaison homologue. (Huertas, 2010)

Un champ d'étude vaste mais encore méconnu est la phosphorylation des différents acteurs de la réparation qui peut influer sur le choix de la voie (figure 11). Par exemple l'activité kinase du complexe DNA-PK inhibe la RH et favorise le NHEJ (Shrivastav et al., 2008). La phosphorylation de CtIP, elle, dépend du cycle cellulaire (Huertas and Jackson, 2009).

Enfin l'environnement chromatinien autour du site de cassure pourrait être un facteur important pour le choix de la voie de réparation des cassures double brin. On sait déjà qu'il y a d'importantes modifications de la chromatine en réponse aux cassures, par exemple la phosphorylation de gamma H2AX qui est un variant de l'histone H2A(Corpet and Almouzni, 2009; Iacovoni et al., 2010) dans la région de la cassure, va aider aux recrutements des

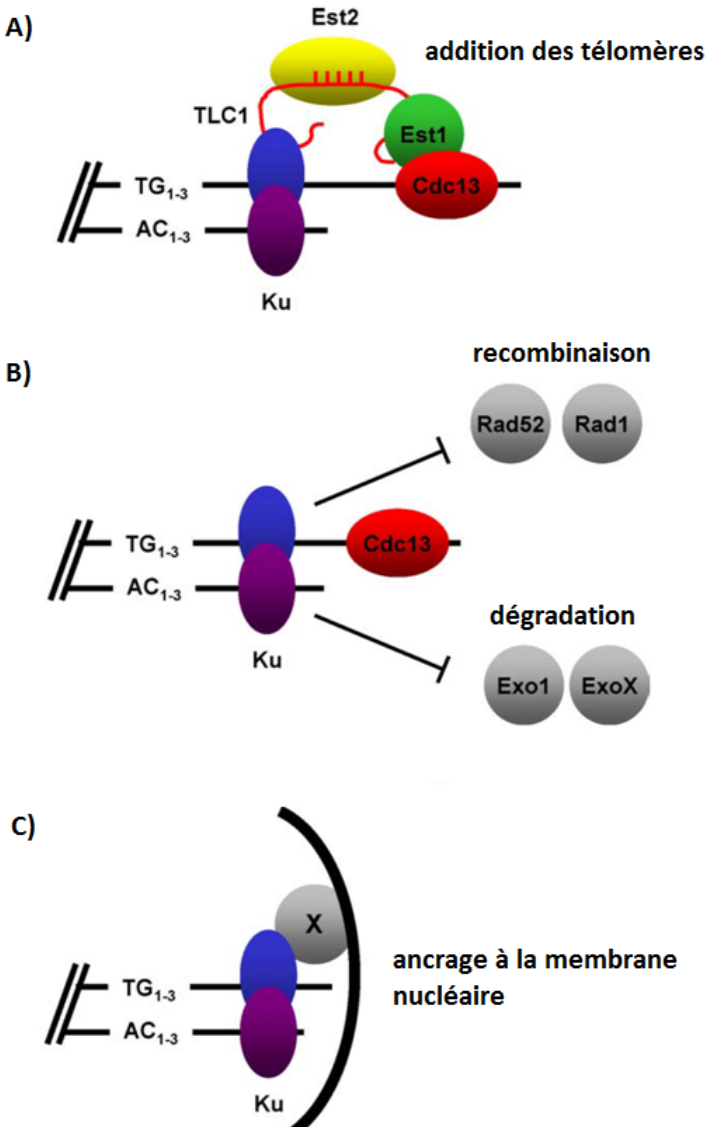


Figure 12 : Les multiples rôles de Ku au niveau des télomères. A) Addition et maintenance des télomères. B) Prévention des phénomènes de translocations chromosomiques et de la dégradation par les nucléases. C) Ancrage des télomères à la membrane nucléaire. (Fisher and Zakian, 2005)

protéines de réparation mais on ne sait pas encore si les marques épigénétiques déjà présentes sur la chromatine avant l'introduction des cassures favorisent l'utilisation d'une voie ou d'une autre. Il a également été observé que l'histone H1 avait tendance à favoriser la réparation par le alt-NHEJ. (Rosidi et al., 2008)

Choisir de ne pas réparer

Le cas des télomères

Les télomères sont des séquences répétées, non codantes, situées aux extrémités des chromosomes linéaires, où elles ont un rôle protecteur. Les télomères permettent de protéger l'ADN des accidents de réplication fréquents aux extrémités de chromosomes. En effet la polymérase n'est pas capable de répliquer l'ADN aux extrémités, c'est pourquoi l'ADN au niveau de ces extrémités est non codant. Un mécanisme spécialisé s'assure que les pertes de séquences sont remplacées par la télomerase. Un prix Nobel en 2009 est venu récompenser le travail d'Elisabeth Blackburn pour son travail sur la télomérase chez le cilié *Tetrahymena thermophila.*

Cependant on a observé aux extrémités des chromosomes la présence de protéines de réparation. Notamment Rad50 et la protéine Ku. Ces observations sont surprenantes car il est indispensable pour la cellule de préserver les chromosomes des translocations. Le recollement des extrémités des chromosomes entrainerait des problèmes de séparation des chromosomes, très dangereux pour la cellule. On aurait donc pu imaginer que les protéines de réparation aient tendance à être exclues de la zone des télomères. Mais des études menées chez la levure ont permis de mettre en évidence que la protéine Ku est capable d'interagir avec la sous unité ARN de la télomérase (figure 12), et est impliquée dans l'addition et la maintenance des télomères (Pfingsten et al., 2012). L'absence de Ku a tendance à favoriser l'apparition de

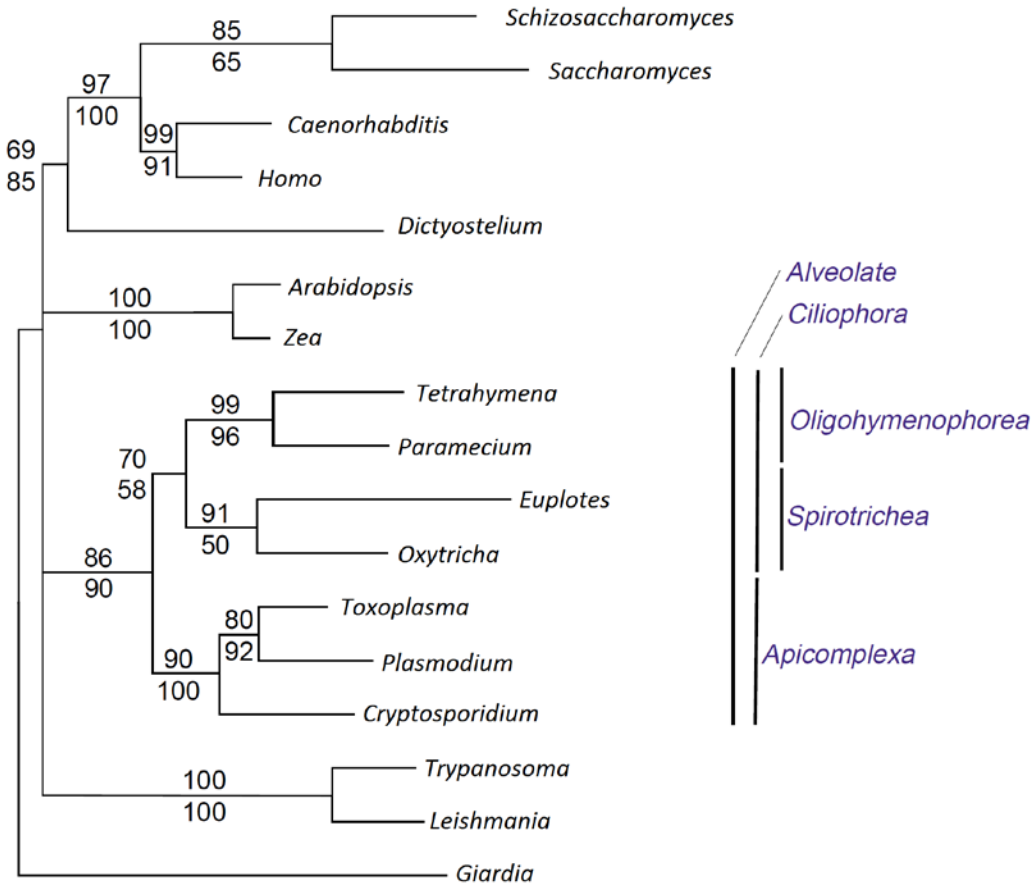


Figure 13 : Arbre des eucaryotes et classification simplifiée des principaux ciliés évoqués dans ce mémoire. On observe que le groupe monophylétique le plus proche est celui des apicomplexes. (Doak et al., 2003)

translocations chromosomiques via alt-NHEJ (Indiviglio and Bertuch, 2009). L'hétérodimère Ku a donc ici un rôle protecteur empêchant l'action de nucléase souvent impliquées dans les premières étapes du alt-NHEJ. Enfin, l'hétérodimère Ku semble avoir un rôle d'ancrage des télomères à la membrane nucléaire via des partenaires encore inconnus (Fisher and Zakian, 2005).

La paramécie comme organisme modèle.

Paramecium tetraurelia est un eucaryote unicellulaire appartenant au groupe des ciliés (figure 13). Comme tous les ciliés, la cellule est couverte d'un grand nombre de cils arrangés régulièrement qui permettent le déplacement par nage, et la nutrition en attirant les nutriments en suspension dans le milieu (bactéries et autres microorganismes) au niveau de la bouche.

Une autre caractéristique majeure des ciliés est la présence d'un dimorphisme nucléaire. En effet, la paramécie possède deux types de noyaux au sein d'un même cytoplasme. Les deux micronoyaux (mic) qui sont diploïdes et le macronoyau (MAC) qui est fortement polyploïde. On estime la ploïdie du MAC entre 800 et 1000n (Berger, 1973).

En fait, la paramécie est un unicellulaire eucaryote qui est parvenu à séparer ses fonctions germinales et somatiques au sein d'un unique cytoplasme. Les micronoyaux peuvent être assimilés à la lignée germinale, ils subissent la méiose et transmettent l'information génétique à la génération sexuelle suivante pendant la conjugaison ou l'autogamie. Le macronoyau constitue la lignée somatique, son génome est réarrangé et il assure l'expression des gènes, il ne transmet par l'information génétique à la génération suivante car il est perdu à chaque cycle sexuel et un nouveau macronoyau doit être généré à chacun de ces cycles. La paramécie constitue un organisme modèle pour l'analyse des fonctions ciliaires mais aussi pour les mécanismes de réarrangements programmés des génomes. Il est très facile d'éteindre

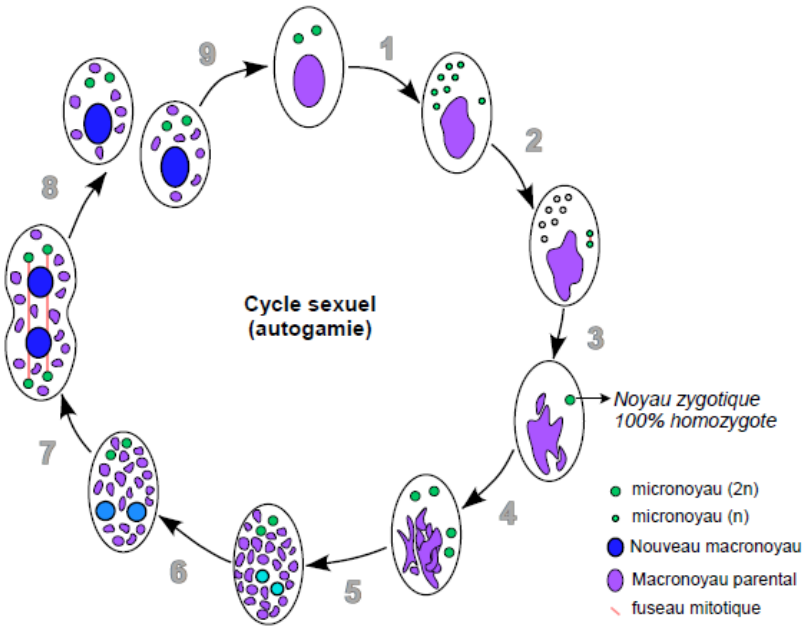


Figure 14 : Représentation schématique du cycle sexuel de l'autogamie. Cela concerne les cellules "âgées" et n'implique pas de partenaire sexuel.

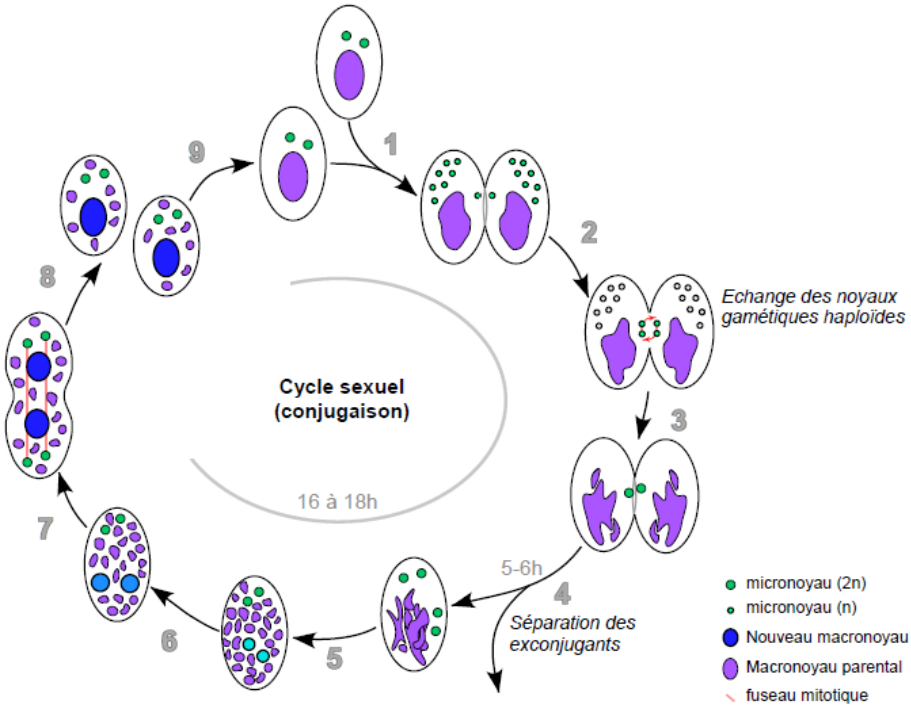


Figure 15 : Représentation schématique du cycle sexuel de conjugaison. Cela concerne des cellules "jeunes" et nécessite un partenaire de type sexuel compatible.

l'expression des gènes par ARN interférence par un mécanisme semblable à celui utilisé chez le nématode *C. elegans.* Les organismes sont nourris avec des bactéries produisant des ARNs double brin qui seront pris en charge pas la machinerie d'ARN interférence de la paramécie qui dégradera l'ARN messager homologue. On peut transformer le macronoyau végétatif en y injectant de l'ADN ; cet ADN sera alors linéarisé s'il ne l'était pas déjà, cet ADN ne s'intégrera pas dans le génome par recombinaison homologue mais des télomères seront ajoutés aux extrémités, ce qui permettra son maintien sous forme de pseudochromosomes jusqu'au prochain cycle sexuel. On peut induire des délétions somatiques dans le macronoyau. Les cycles sexuels peuvent être induits par carence alimentaire.

La paramécie, sa vie.

Vie végétative

La paramécie se propage en phase végétative par division. Les micronoyaux se divisent par mitose classique, tandis que le macronoyau fait une division dite « amitotique », il n'y a pas condensation des chromosomes. En effet, pendant la division le macronoyau va se séparer en deux lots d'ADN à peu près équivalent en s'étirant puis finissant par se séparer. L'ADN est ensuite amplifié pour atteindre un niveau de ploïdie « final » d'environ 800n.

Processus sexuels

La paramécie possède deux types de reproduction sexuée, l'autogamie (figure 14) et la conjugaison (figure 15) Lorsque la paramécie est dite « vieille », ce qui signifie qu'elle a accumulé plus d'une vingtaine de divisions végétatives, et qu'elle est en condition de carence alimentaire, l'autogamie peut se déclencher. L'autogamie débute par la méiose des deux micronoyaux, ce qui produit huit produits de méiose haploïdes. En même temps, le

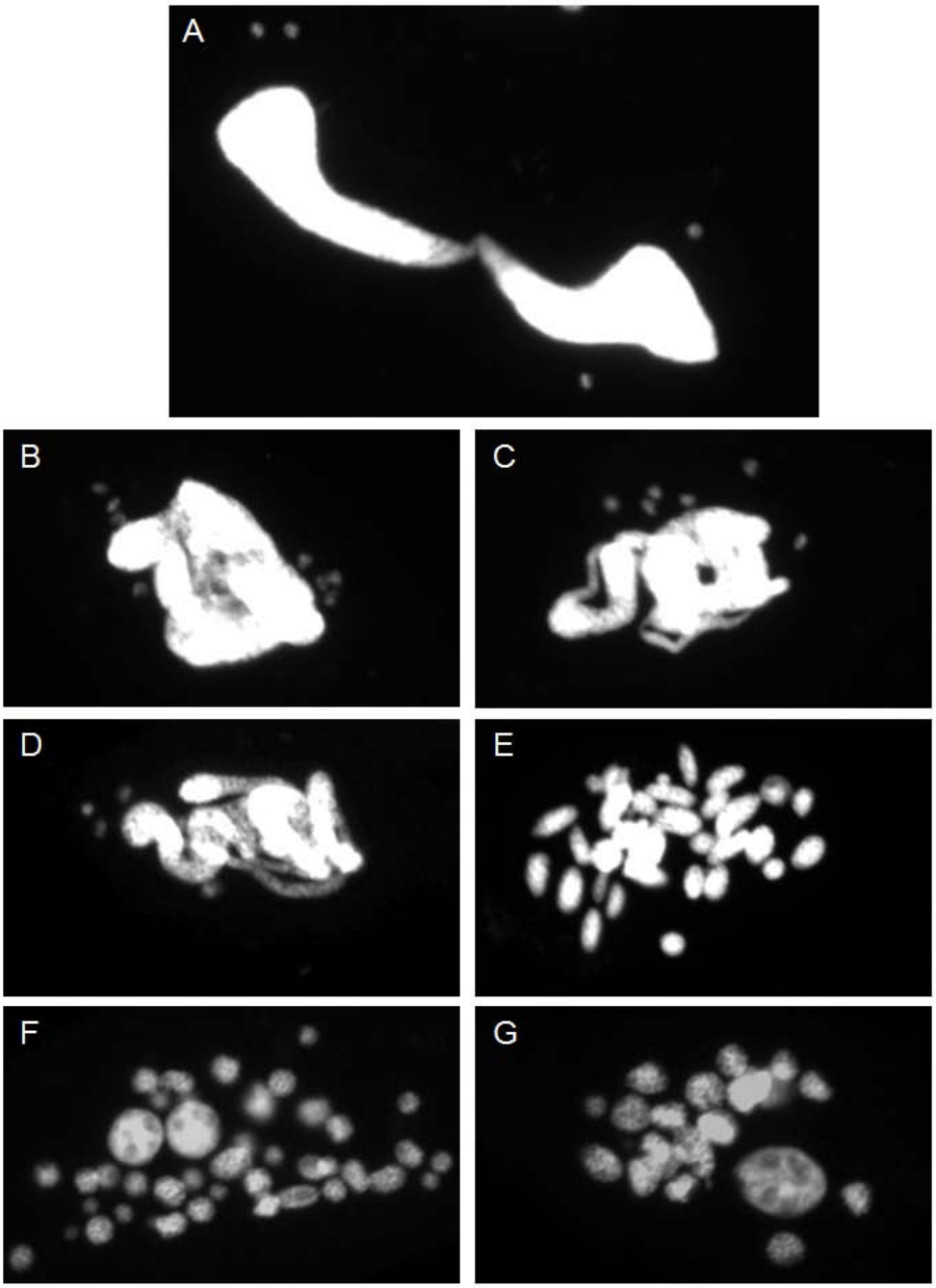


Figure 16 : Marquage au DAPI des noyaux de cellules végétatives et autogames. A) Cellule végétative en division, on observe que le MAC fait une division amitotique. B) Méiose. C) Méiose et « débobinage » de l'ancien MAC. D) Débobinage de l'ancien MAC. E) Ancien MAC fragmenté. F) Ancien MAC fragmenté et deux nouveau MAC en développement. G) Caryonide, les deux MACs en développement ont été ségrégés pendant une division, il reste toujours des fragments de l'ancien MAC qui vont progressivement disparaitre.

macronoyau commence à se fragmenter. Sept des produits de méiose vont dégénérer tandis que le dernier se divise pour donner deux noyaux haploïdes identiques, les noyaux gamétiques. Ceux-ci fusionnent pour former un noyau diploïde, le noyau zygotique. Celui-ci va subir deux mitoses successives pour obtenir quatre copies du noyau zygotique. Deux d'entre eux vont migrer à l'avant de la cellule et les deux autres à l'arrière. Les deux à l'avant vont se différencier en micronoyaux tandis que les deux autres vont se différencier en macronoyaux (figure 16). Une division dite « caryonidale » va avoir lieu ; au cours de celle-ci les micronoyaux vont faire une mitose classique tandis que les macronoyaux toujours en développement vont être ségrégés dans chacune des deux cellules filles. Pendant ce processus les fragments de l'ancien macronoyau sont toujours transcrits et assurent la vie de la cellule en attendant que les nouveaux macronoyaux prennent progressivement le relais. L'autogamie, de par son mécanisme, génère une descendance homozygote.

L'autre mode de reproduction sexuée est la conjugaison, également induite par carence alimentaire mais impliquant des cellules ayant accumulé moins de divisions végétatives; elles sont dites « jeunes ». La conjugaison se déroule entre deux partenaires de types sexuels compatibles. Les mécanismes sont très semblables à ceux impliqués dans l'autogamie, la différence majeure est qu'il y a échange d'un noyau gamétique entre les deux partenaires. Les deux noyaux gamétiques, un de chaque partenaire vont alors fusionner pour former le noyau zygotique. Le reste du cycle se déroule normalement comme lors de l'autogamie.

La paramécie, son œuvre.

Pendant ces processus et à chaque cycle sexuel, le macronoyau est perdu et la paramécie doit produire un nouveau MAC à partir d'un noyau zygotique. Les récents progrès en matière de séquençage haut débit ont permis de séquencer le génome macronucléaire il y a quelques années (Aury et al., 2006), tandis que le séquençage du génome micronucléaire, toujours en

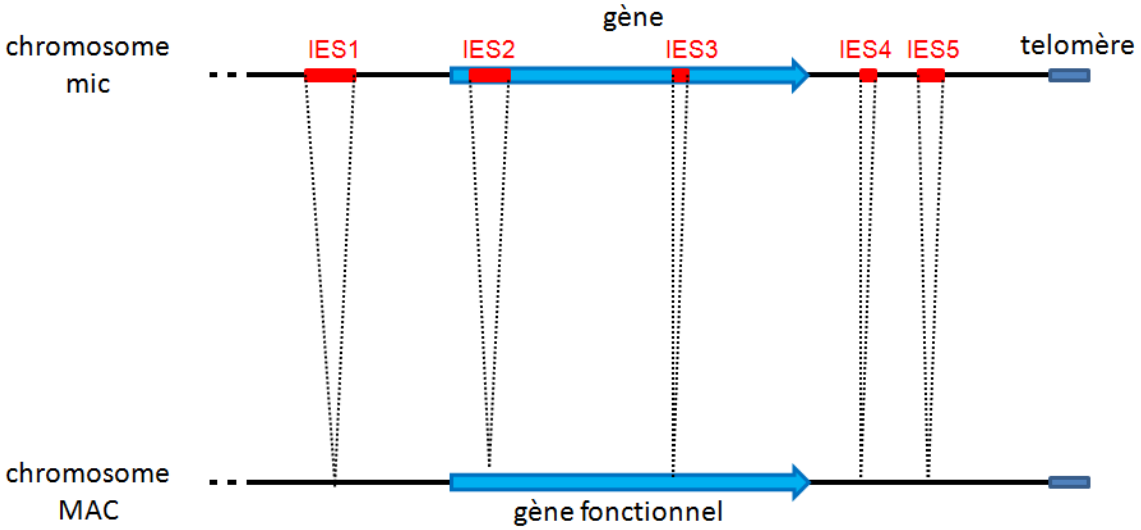


Figure 17 : Elimination des IES pouvant se trouver dans des régions intergéniques ou interrompre des cadres ouverts de lecture. Les IES sont éliminées précisément au nucléotide près.

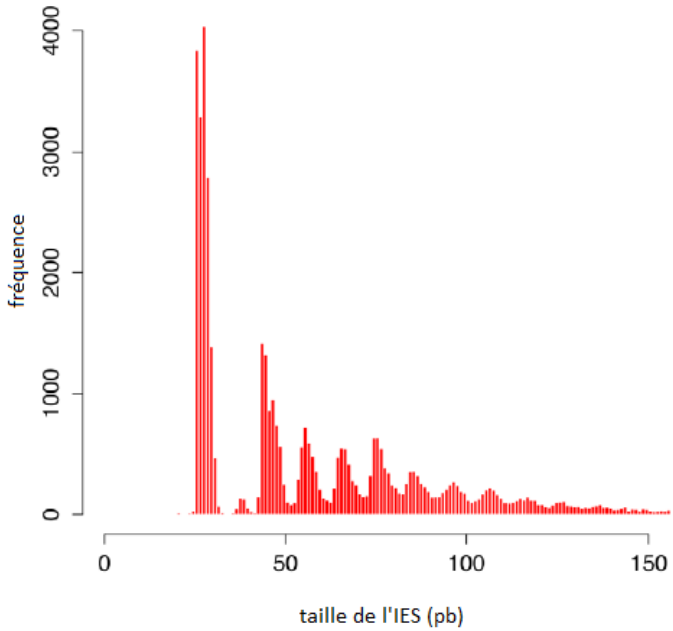


Figure 18 : Distribution de la taille des IES inférieure à 150 pb. On remarque une périodicité de 10-11 pb. Le pic autour de 26 pb représente un tiers des IES. Les IES autour de 36 pb semblent contreselectionnées. (Arnaiz et al., 2012)

cours d'assemblage et d'annotation, n'a été réalisé qu'il y a quelques mois (Arnaiz et al., 2012). En effet la très forte ploïdie du macronoyau rendait difficile le séquençage du micronoyau. La comparaison des génomes a pu confirmer qu'il existe des différences significatives entre les génomes micronucléaire et macronucléaire (Baudry et al., 2009). Ainsi pour passer d'un génome micronucléaire à un nouveau génome macronucléaire, plusieurs processus simultanés de réarrangements programmés du génome sont nécessaires: l'amplification de l'ADN pour atteindre le niveau de ploïdie final d'environ 800n ; l'excision précise de courtes séquences appelées IES; l'élimination hétérogène de séquences répétées.

Les réarrangements programmés du génome.

L'excision précise des IES

Un jeu de référence de 45 000 IES a été décrit récemment par comparaison des génomes mic et MAC. Les IES, pour Internal Eliminated Sequences sont des séquences uniques et non codantes, pouvant faire entre plusieurs dizaines et plusieurs milliers de paires de base, bien que plus de 90 % aient une taille inférieure à 150 pb. Ces séquences peuvent se trouver dans les régions intergéniques mais tout aussi bien interrompre les cadres ouverts de lecture. Le séquençage a montré que 47% des gènes sont interrompus par au moins une IES (Arnaiz et al., 2012). Pour rétablir des gènes fonctionnels dans le génome MAC, il est donc indispensable que l'excision des IES se fasse de façon extrêmement précise, au nucléotide près (figure 17).

Bien que chaque IES soit unique, elles partagent tout de même des caractéristiques communes. Il avait été proposé sur la base de l'analyse de la séquence de quelques dizaines d'IES, qu'elles dérivaient de transposons. En effet un faible consensus avec un transposon de

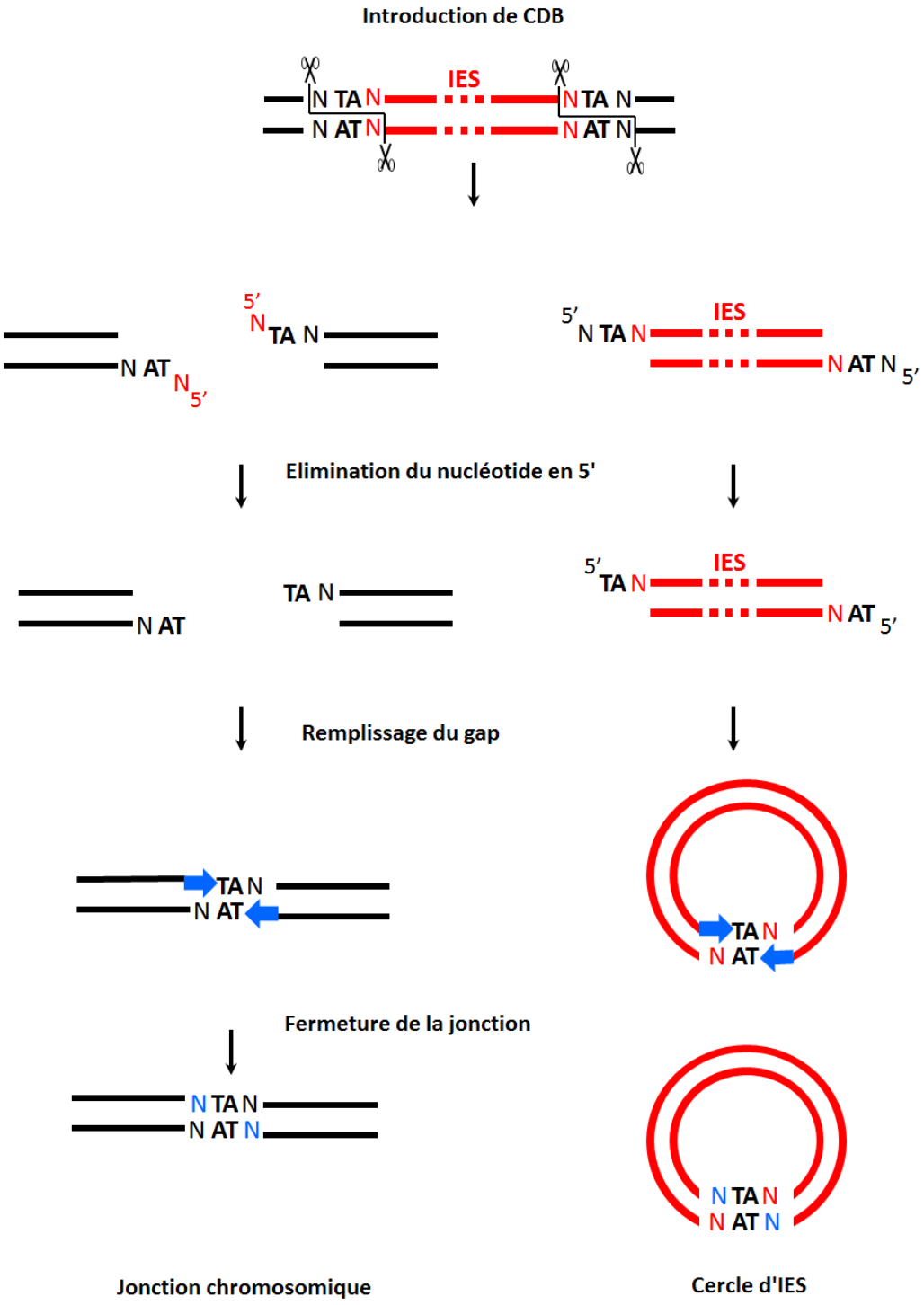


Figure 19 : Modèle d'excision des IES. Coté jonction chromosomique, et coté jonction d'IES. On imagine que les mécanismes sont les mêmes coté IES (Dubois et al., 2012).

type Tc1/mariner a pu être décrit au bornes des IES, avec un dinucléotide TA absolument conservé à toutes les bornes d'IES. La cicatrice d'excision laisse un unique TA dans le génome macronucléaire au niveau de la jonction MAC (Gratias and Betermier, 2001). La taille des IES semble avoir été contrainte au cours de l'évolution. On peut voir dans la distribution des tailles d'IES une périodicité d'environ 10 pb, soit la longueur d'un tour d'hélice d'ADN, suggérant une contrainte topologique lors de l'excision (figure 18).

L'excision des IES se fait via l'introduction concertée des cassures double brin de l'ADN à chaque borne des IES. Les cassures introduites ont une géométrie bien spécifique de 4 bases sortantes en 5' centré sur le dinucléotide TA. Les extrémités sortantes des séquences MAC flanquantes ne sont pas toujours directement compatibles. C'est pourquoi le modèle propose des étapes de maturation des extrémités. Le dernier nucléotide en 5' est enlevé puis le gap ainsi formé au niveau de la synapse est comblé par une activité de polymérisation 3' vers 5' avant réparation (figure 19).

PiggyMac, transposase domestiquée

Les travaux de Céline Baudry et Sophie Malinsky publiés en 2009 (Baudry et al., 2009) ont permis de mettre en évidence l'action de PiggyMac (Pgm), une transposase domestiquée de la famille des transposases *piggybac* lors des réarrangements programmés du génome. Le gène de cette transposase a été intégré dans le génome de la paramécie, a divergé du gène de transposase *piggybac* de référence mais a conservé les principaux domaines ainsi que les résidus catalytiques importants pour l'activité de coupure de l'ADN, la triade DDD. Elle possède des domaines additionnels en Nter et Cter (figure 20) dont la caractérisation est en cours au laboratoire. Lorsque ce gène est éteint par ARN interférence, il n'est plus possible de détecter de cassures double brin programmées aux bornes des IES, ce qui entraine leur rétention dans le nouveau MAC. Cela n'empêche toutefois pas l'amplification de l'ADN non réarrangé.

Figure 20 : Les gènes PGM et PGM-like de *Paramecium tetraurelia*. Les différents domaines sont indiqués, particulièrement le domaine catalytique avec la triade DDD, le domaine de liaison à l'ADN en bleu, et le domaine riche en cystéines en violet.

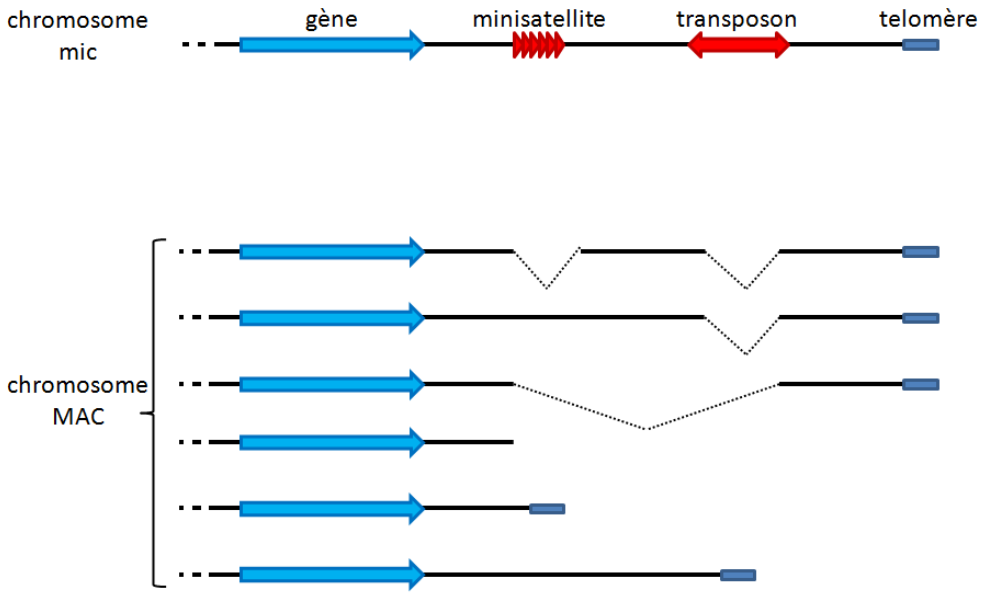


Figure 21 : Représentation schématique de l'élimination hétérogène des séquences répétées. Elle peut mener à des délétions de taille variable ou à la fragmentation des chromosomes avec addition de télomères aux extrémités. Présence de microhomologies au niveau des jonctions.

Ainsi pendant l'autogamie, l'ADN de cellules déplétée pour PiggyMac a été extrait et purifié. C'est par cette technique qu'ont pu être identifié 45 000 IES du premier jeu de référence. Récemment ont été identifié des gènes codant des protéines homologues à Pgm. 9 gènes différents ont été identifiés (figure 20). Les protéines qu'ils codent partagent les domaines cœur de la protéine Piggybac comme Pgm, mais ont plus ou moins divergé. Certains possèdent des domaines additionnels et sont relativement proches de Pgm. Toutes les protéines codées par ces gènes ont leur domaine catalytique partiellement ou complètement muté au niveau de la triade catalytique DDD. Cependant, de façon surprenante, tous ces gènes sont nécessaires pour la survie de la descendance et leur extinction affecte l'excision des IES. L'étude de ces gènes est en cours (J. Bischerour, E. Dubois, M. Nowacki).

L'élimination hétérogène de séquences répétées

Comme précisé plus haut, l'autre type de réarrangement programmé du génome est l'élimination d'éléments répétés. Il s'agit souvent de transposons et de séquences minisatellites. Cela peut conduire à l'élimination de longues séquences d'ADN, de l'ordre de plusieurs dizaines de kb, générant des délétions de tailles variables, soit la fragmentation des chromosomes avec addition de télomères aux extrémités cassées (figure 21) (Le Mouel et al., 2003). Au niveau des jonctions de réparation on observe une forte densité en nucléotide T et A et l'utilisation de zones de microhomologies (Garnier et al., 2004). On connait peu de sites naturels de réarrangements imprécis, en effet, la grande hétérogénéité observée au niveau de ces séquences éliminées est telle qu'il est difficile de décortiquer les mécanismes moléculaires mis en jeu. On peut suivre le devenir des délétions somatiques qui peuvent se transmettre, bien qu'imparfaitement, à la génération suivante.

La protéine Pgm est également nécessaire pour l'élimination hétérogène, en son absence on n'observe pas de fragmentation des chromosomes et les transposons sont retenus. Bien que

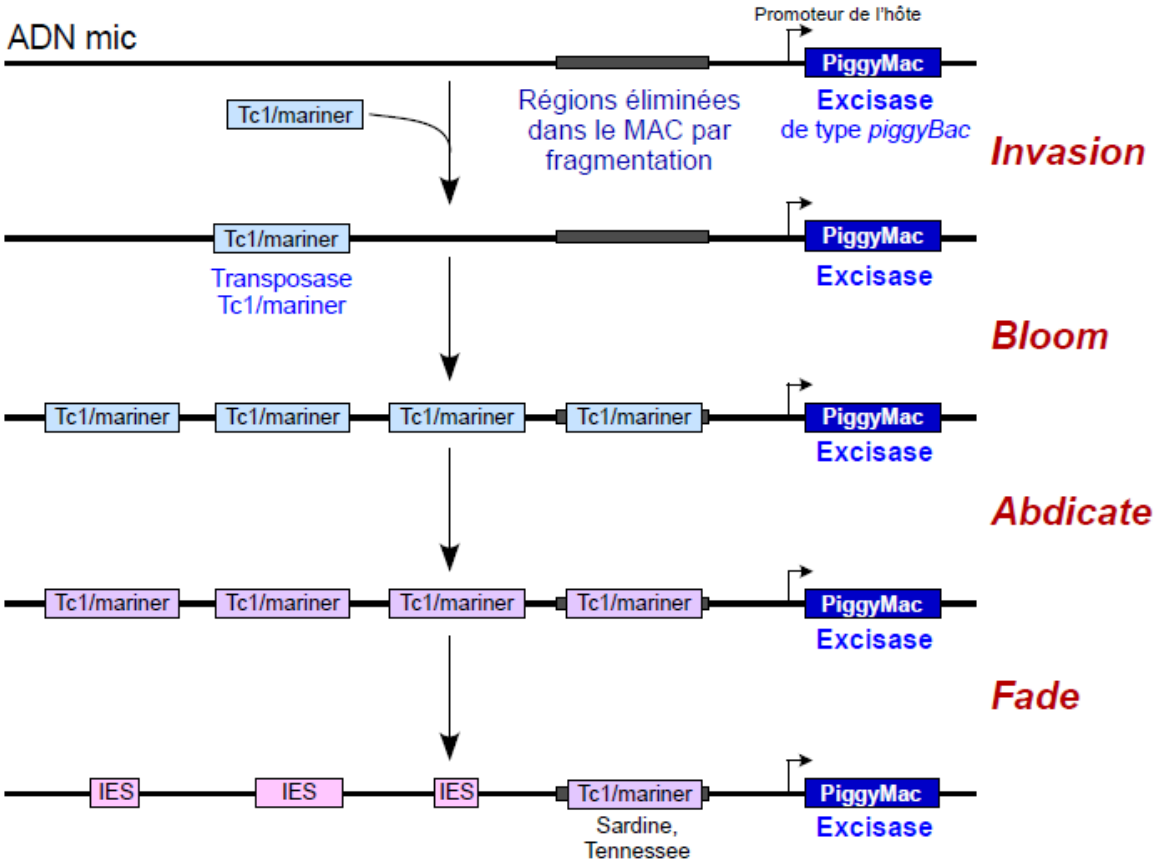


Figure 22 : Le modèle IBAF revisité. La protéine PiggyMac domestiquée apparait ici avant l'invasion du génome par les éléments Tc1/mariner. (Baudry et al., 2009; Klobutcher and Herrick, 1997)

l’on ne sache pas s’il s’agit d’un effet direct ou non. Peut-être faut-il exciser les IES d’un gène essentiel pour les réarrangements imprécis?

Origine des IES et des séquences à éliminer

En 1995, puis en 1997, Klobutcher et Herrick (Klobutcher and Herrick, 1995, 1997) ont proposé un modèle pour expliquer l’origine évolutive des IES à TA. Ce scénario en plusieurs étapes a été appelé IBAF (figure 22) pour Invasion, Bloom, Abdicate, Fade. L’hypothèse centrale de ce modèle est que les IES dérivent de transposons qui ont perdu leur capacité à coder leur transposase mais qui peuvent toujours être excisés.

- Dans la première étape, **Invasion,** le génome est envahi par un élément transposable de type Tc1/mariner, peut être par transfert horizontal. En effet les ciliés ingèrent des bactéries et d’autres eucaryotes unicellulaires qui sont digéré au milieu de la cellule dans les vacuoles digestives. Ce mode d’alimentation peut favoriser le passage des éléments mobiles d’un organisme à l’autre.

- **Bloom** correspond à l’étape dans laquelle les transposons Tc1/mariner actifs envahissent massivement le génome mic. Le dimorphisme nucléaire et le fait que la transcription est assurée par le génome MAC permettent l’envahissement du génome mic, tant que les transposons sont excisés précisément pendant les processus sexuels. Cette excision pourrait être assurée par leur propre excisase, mais Klobutcher et Herrick proposaient déjà qu’une autre transposase ou nucléase puisse assurer l’excision précise.

- **Abdicate,** correspond à la domestication de la transposase des éléments Tc1/mariner. Permettant à la machinerie cellulaire de faire passer l'excision des éléments transposables, pendant la différenciation macronucléaire, sous son contrôle.

- Enfin, dans l'étape **Fade**, les transposons Tc1/mariner, libérés de la pression sur leur capacité à coder leur propre transposase mais pas de leur capacité à être excisés, auraient pu dégénérer tout en conservant leurs bornes permettant l'excision.

La découverte de la protéine Pgm a amené à reconsidérer ce modèle. En effet si on imagine qu'une transposase piggybac était présente avant même l'invasion du génome par les éléments Tc1/mariner, puis domestiquée pour prendre en charge leur élimination du génome macronucléaire, l'étape Abdicate est inutile. Puisque Pgm est présente, les Tc1/mariner sont déjà libérés de la pression de sélection sur leur capacité codante et peuvent dégénérer si les bornes permettant l'excision sont conservées.

Bien sur, il est tout à fait possible qu'il y ait eu plusieurs vagues d'invasion par les transposons au cours de l'évolution et qu'ils aient dégénéré à des vitesses différentes. Cela expliquerait la présence actuelle dans le génome des IES et de transposons Tc/mariner plus ou moins conservés, et probablement d'autres encore à découvrir dans les régions éliminées de façon imprécise.

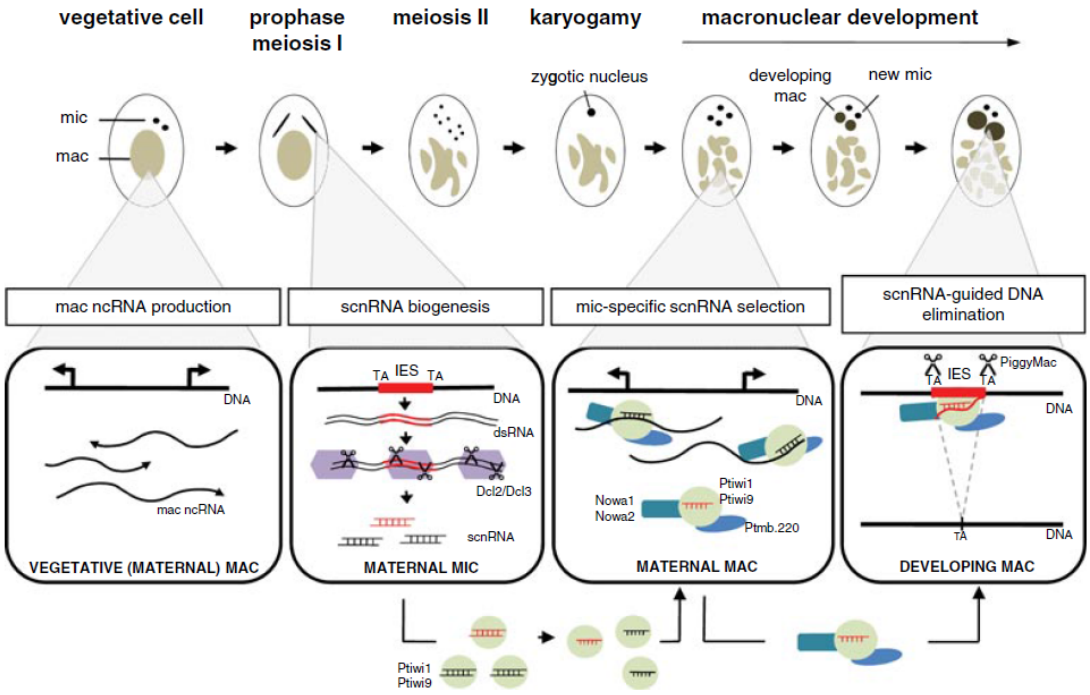


Figure 23 : Modèle de contrôle des réarrangements programmés du génome par les scanARNs (Coyne et al., 2012).

Contrôle des réarrangements programmés du génome.

Dans les deux cas, réarrangements précis et imprécis, la question de savoir comment la cellule fait pour reconnaitre les régions à éliminer se pose. En effet, il n'y a pas de séquence spécifique conservée qui pourrait être le site de fixation de Pgm. Le modèle proposé repose sur la comparaison des génomes du mic et de l'ancien MAC via des intermédiaires ARN, on appelle ce mécanisme le « scanning » (Duharcourt et al., 2009). De façon schématique, les ARNs non codants permettant de cibler les séquences à éliminer seraient sélectionnés puis serviraient à la dépose, sur le génome à réarranger, de marques épigénétiques qui permettrait le ciblage par la ou les machineries d'élimination (figure 23) . C'est un mécanisme proche de ce qui est observé pour l'extinction transcriptionnelle des transposons et autres éléments mobiles, neutralisés par des marques épigénétiques. Chez la paramécie, la cellule va une étape plus loin en éliminant physiquement ces séquences. Voici les différentes étapes du processus de scanning et d'élimination :

- Des longs transcrits non codants sont produits par l'ancien MAC de façon constitutive et représentent une copie ARN du génome réarrangé chez le parent (Lepere et al., 2008)
- Pendant la méiose, des transcrits non codants de 25 nt (scnARN) sont produits par les mics, de façon dépendante de Dicer 2 et 3 (Lepere et al., 2009). Ces ARNs pourraient couvrir l'ensemble du génome non réarrangé du mic.
- Ces scnARN vont être transportés, de façon dépendante de Ptiwi 1 et 9 jusqu'à l'ancien MAC où aura lieu la comparaison des versions germinales et réarrangés des génomes parentaux. Les scnARN qui s'apparient avec les transcrits de l'ancien MAC réarrangé vont être retenus et les autres, correspondant aux régions éliminées dans l'ancien MAC et à éliminer dans le nouveau, vont être transportés de façon dépendante de Ptiwi 1 et 9 ainsi que de Nowa 1 et 2 jusqu'au MAC en développement.

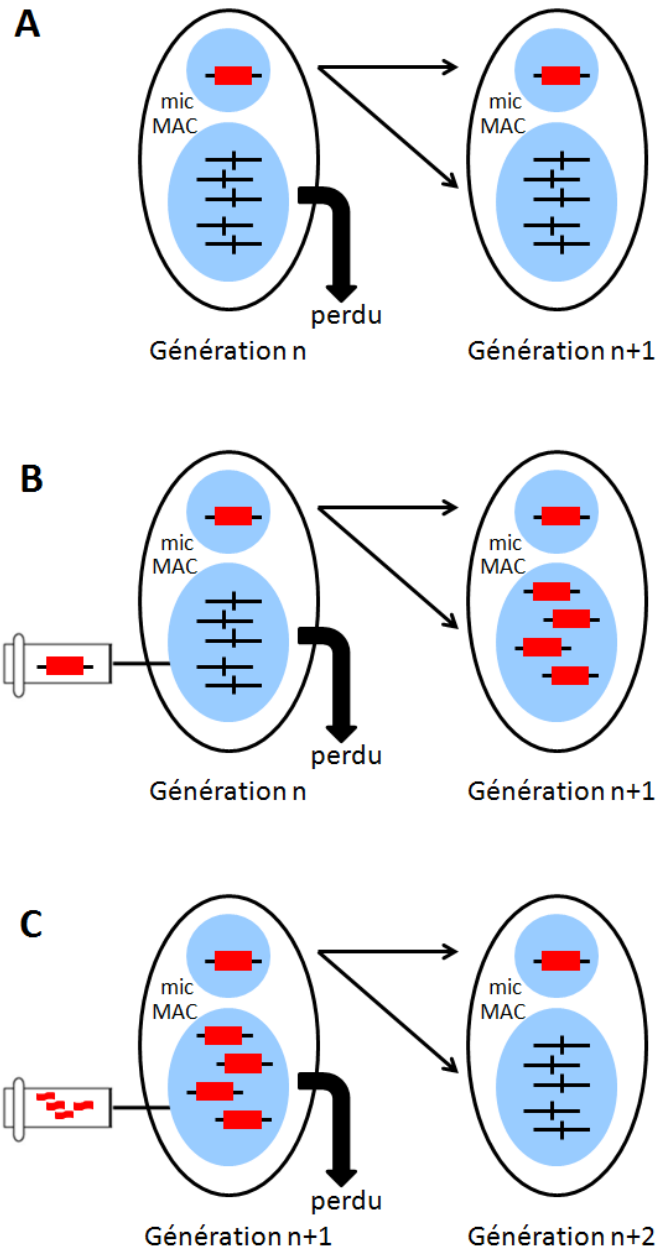


Figure 24 : Contrôle épigénétique des réarrangements. A) situation normale dans une souche sauvage. B) L'ancien MAC est transformé avec l'IES rouge, cette IES est maintenue dans la descendance. C) L'injection de petits ARNs ciblant et dégradant le long transcrit non codant produit dans l'ancien MAC restaure l'excision des cette IES dans la nouvelle génération. (Duharcourt et al., 1995; Duharcourt et al., 1998).

- On imagine que les scnARN vont s'apparier aux transcrits naissants produits par les nouveaux MACs en développement de façon dépendante d'un facteur de transcription TFIIS spécialisé (K. Maliszewska, communication personnelle), cela permettant de cibler les séquences à éliminer par l'ajout des marques épigénétiques encore non identifiées sur la chromatine ou sur l'ADN.
- La transposase domestiquée Pgm est recrutée au niveau de ces marques épigénétiques et les séquences spécifiques de la lignée germinale éliminées.

Les études expérimentales sur plusieurs IES (figure 24) , avant le séquençage haut débit du MAC, ont permis de montrer que ce modèle de contrôle maternel de l'excision par les scnARN pouvait rendre compte des propriétés d'au moins une partie des IES, en plus d'identifier certains des acteurs protéiques impliqués dans ce processus, tel que Dicer (Lepere et al., 2009), Ptiwi (Bouhouche et al., 2011), et Nowa (Nowacki et al., 2005).

Ainsi :

- L'injection dans le MAC parental d'un ADN contenant une IES induit dans certains cas la rétention de cette IES dans le nouveau MAC de la nouvelle génération. Cette IES peut ensuite être maintenue dans les générations suivantes (Duharcourt et al., 1995; Duharcourt et al., 1998).
- Le ciblage et la dégradation du long transcrit non codant produit dans l'ancien MAC, dans la souche qui n'excise pas l'IES, restaurent l'excision de cette IES dans la nouvelle génération (Lepere et al., 2008).

Cependant ces résultats ne sont pas valables pour toutes les IES. On observe un gradient d'efficacité de rétention. Seules deux IES sur quatorze testées sont retenues à 100% après injection d'ADN contenant l'IES dans le MAC parental, certaines présentent un niveau de rétention intermédiaire, d'autres ne sont pas du tout affectées. (Duharcourt et al., 2009).

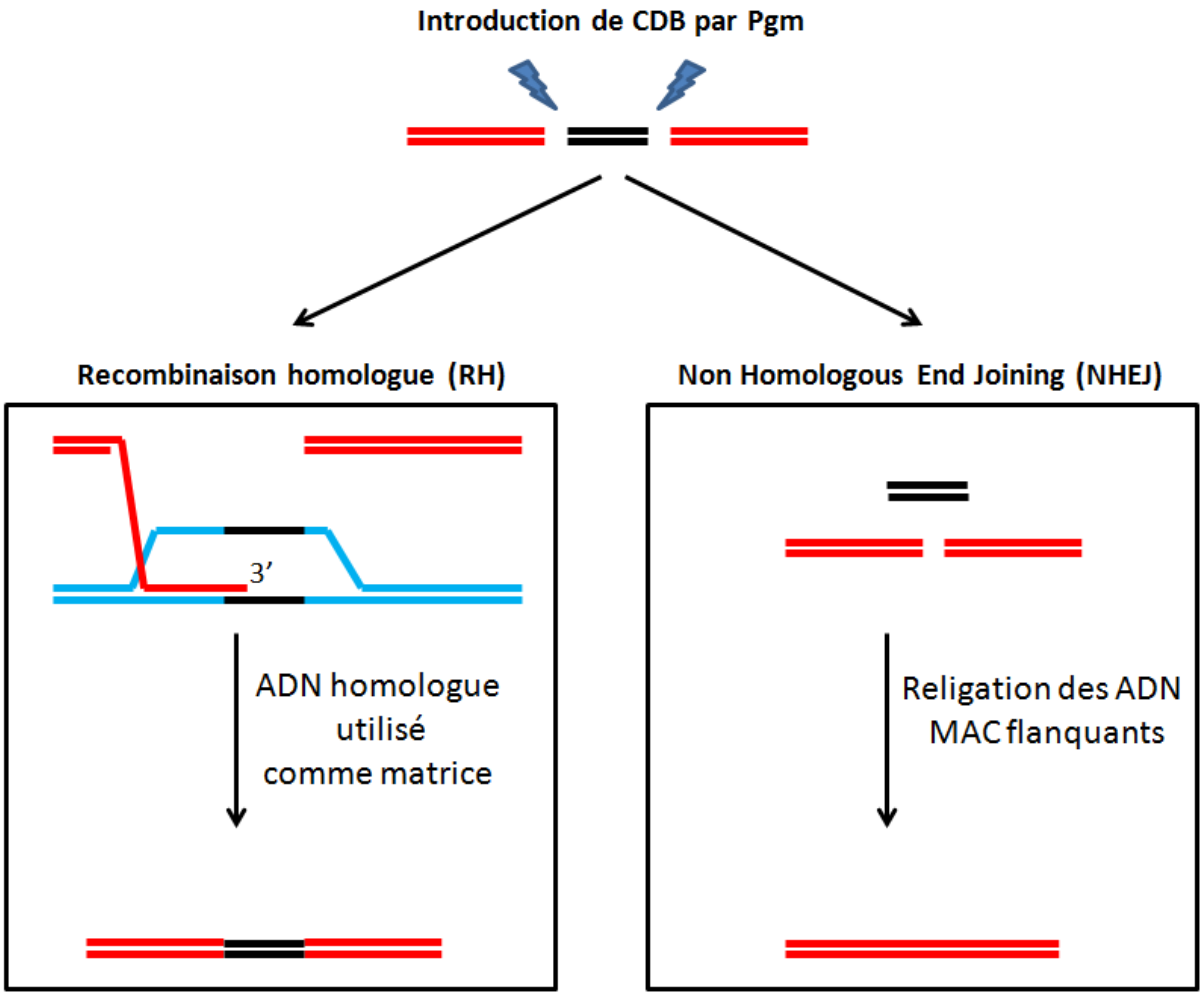


Figure 25 : Comment réparer les cassures introduites par Pgm? Représentation schématique des voies de réparation possible lors de l'excision des IES (ici en noir).

Quelle voie choisir pour réparer les cassures double brin programmées?

Nous savons qu'il existe au moins 45000 IES, ce qui signifie que rien que pour éliminer les IES, pendant les réarrangements programmés du génome, au moins 90 000 cassures double brin sont introduites et réparées. Cela signifie que la cellule est capable de réparer avec une extrême précision des cassures double brin, réparties tous les un à deux kb le long du génome, tout en évitant les événements de translocation de chromosomes. Il faut aussi se rappeler que l'amplification du génome pour atteindre la ploïdie de 800n et l'élimination des séquences germinales se chevauchent largement. La réplication ne pouvant traverser une cassure double brin de l'ADN, Il est indispensable pour la cellule d'avoir un système de réparation des cassures double brin efficace et précis pour surmonter ce défi moléculaire. Pendant les processus sexuels, de nombreux noyaux sont présents dans la cellule selon les moments ; fragments de l'ancien MAC, produits de méiose haploïdes, diploïdes, nouveaux mic et nouveau MAC en développement. Cependant, bien qu'il existe des échanges d'informations entre ces noyaux, via les ARN non codants notamment, ils conservent leur intégrité, sans rupture des enveloppes nucléaires. Ainsi chacun des noyaux est un lot d'ADN individualisé.

Pendant les réarrangements programmés du génome, dans les nouveaux MAC en développement, il n'y a pas d'ADN qui soit préalablement réarrangé, l'ADN qui y est amplifié provient du noyau zygotique, toutes les molécules d'ADN sont donc issues de la lignée germinale et donc non réarrangées au départ. Ainsi pour réparer ces cassures double brin tout en autorisant l'élimination des IES, il est impossible d'utiliser la recombinaison homologue, car il n'existe pas de matrice déjà réarrangée dans le MAC en développement (figure 25). L'utilisation de cette voie, bien que (ou justement « parce que ») très précise et fidèle ne pourrait mener qu'au maintien des séquences à éliminer en supposant que toutes les

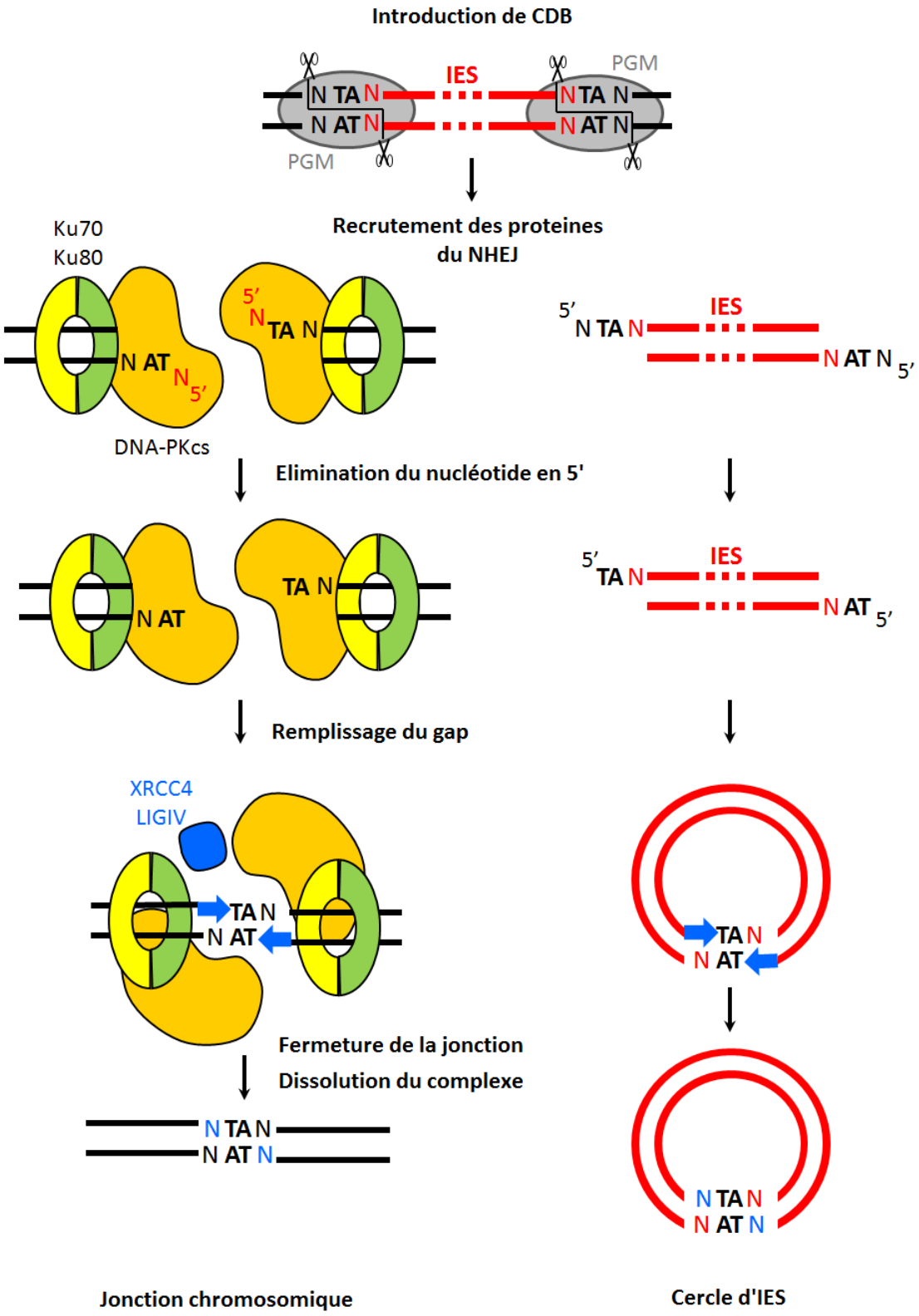


Figure 26 : Modèle de réparation des jonctions d'excision des IES. Coté jonction chromosomique, et coté jonction d'IES. On imagine que les mécanismes sont les mêmes coté IES (Dubois et al., 2012).

molécules ne soient pas amplifiées en même temps. C'est pourquoi le laboratoire s'est intéressé à la voie du NHEJ. En effet cette voie ne nécessite pas de matrice homologue pour la réparation et si les circonstances s'y prêtent, peut être extrêmement précise. Il « suffit » d'introduire les cassures, éliminer les IES et recoller les extrémités flanquantes.

L'excision des IES et la voie NHEJ

Les travaux effectués au laboratoire par Aurélie Kapusta ont démontré que les protéines de paramécie homologues de XRCC4 et LigIV sont absolument nécessaires pour les réarrangements (Kapusta et al., 2011). En effet lorsque ces protéines sont déplétées pendant l'autogamie, les cassures sont introduites normalement aux bornes des IES mais ne sont pas réparées et s'accumulent sans être dégradées. L'absence de réparation de milliers de cassures double brin tout le long du génome empêche l'amplification de l'ADN ; cela est visualisable par marquage de l'ADN au DAPI, on observe un défaut d'amplification d'ADN dans les MAC en développement dans des conditions d'ARNi pour LigIV ou XRCC4. Ainsi des protéines spécifiques de la voie NHEJ sont indispensables pour la réparation des cassures double brin programmées. Cela suggère fortement, même si ce n'est pas encore prouvé, l'implication de l'ensemble de la voie NHEJ dans la réparation des jonctions d'excision des IES.

La figure (figure 26) ci contre présente le modèle d'excision des IES avec l'utilisation des protéines de la voie NHEJ.

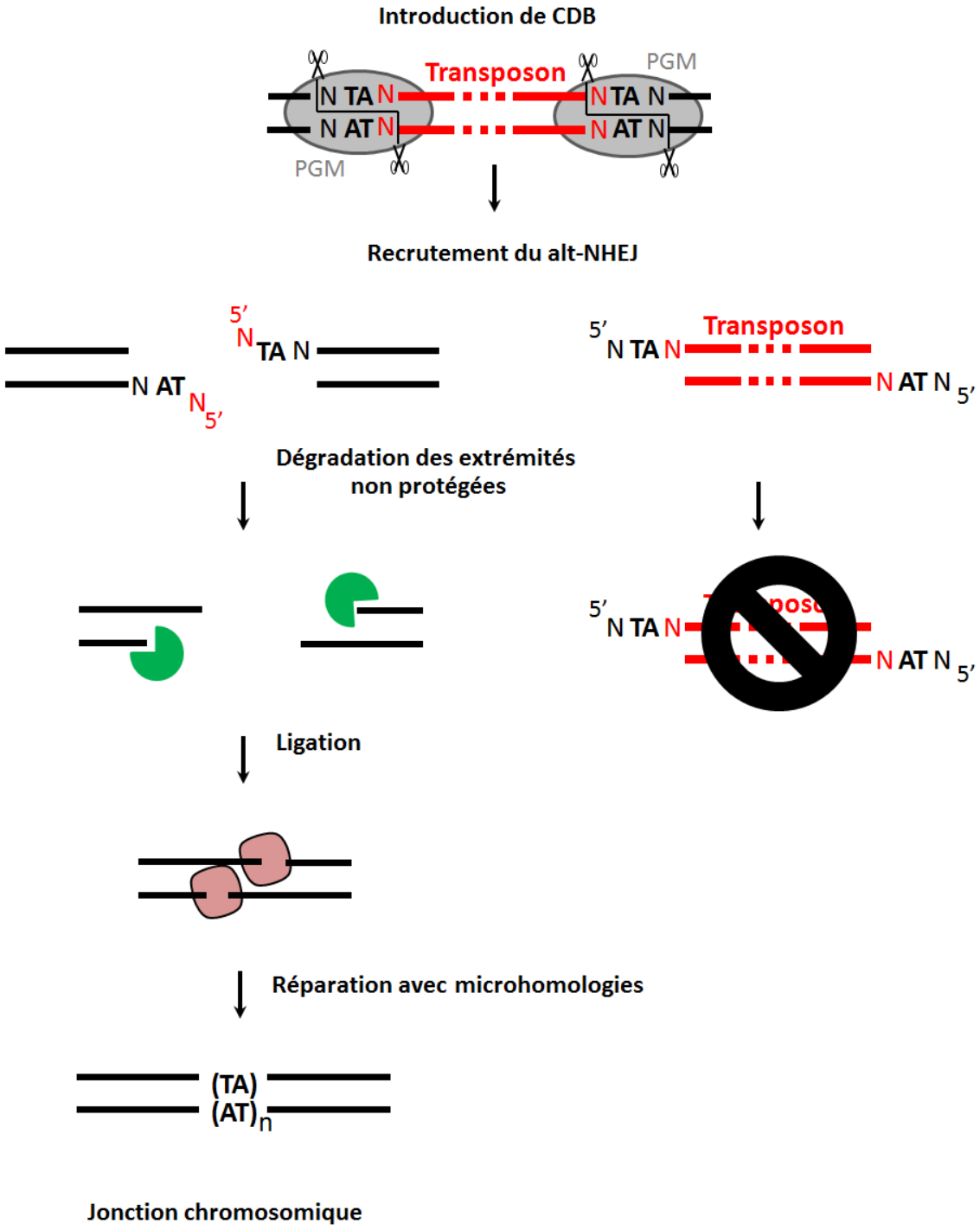


Figure 27 : Modèle de réparation des jonctions d'élimination hétérogène des éléments répétés par la voie alt-NHEJ.

Elimination des séquences répétées …

… et la voie du alt-NHEJ ?

Le mécanisme gouvernant l'élimination hétérogène des séquences répétées est encore inconnu mais considérant que Pgm est nécessaire pour leur élimination, il est possible d'émettre des hypothèses.

La présence d'une forte densité en nucléotides TA, de microhomologie et d'une forte hétérogénéité au niveau des jonctions des délétions intrachromosomiques suggère une réparation par la voie du alt-NHEJ. En effet cette voie est imprécise et est caractérisée par l'usage de microhomologie au niveau de la jonction.

On imagine un modèle où les cassures sont introduites par la protéine Pgm autour des transposons ou des minisatellites de la même manière qu'aux bornes des IES. Les extrémités ne sont pas protégées par l'hétérodimère Ku70/Ku80 et accessibles aux nucléases. Les extrémités ne sont donc plus directement religables. L'étape de ligation finale implique donc l'utilisation de microhomologie (figure 27).

Un modèle alternatif pourrait être que l'hétérogénéité observé au niveau des jonctions soit due non pas à l'imprécision de la réparation mais à une hétérogénéité des sites de coupures par la protéine Pgm. C'est cette situation qui est observée lors de l'excision des IES chez le cilié *Tetrahymena thermophila.*

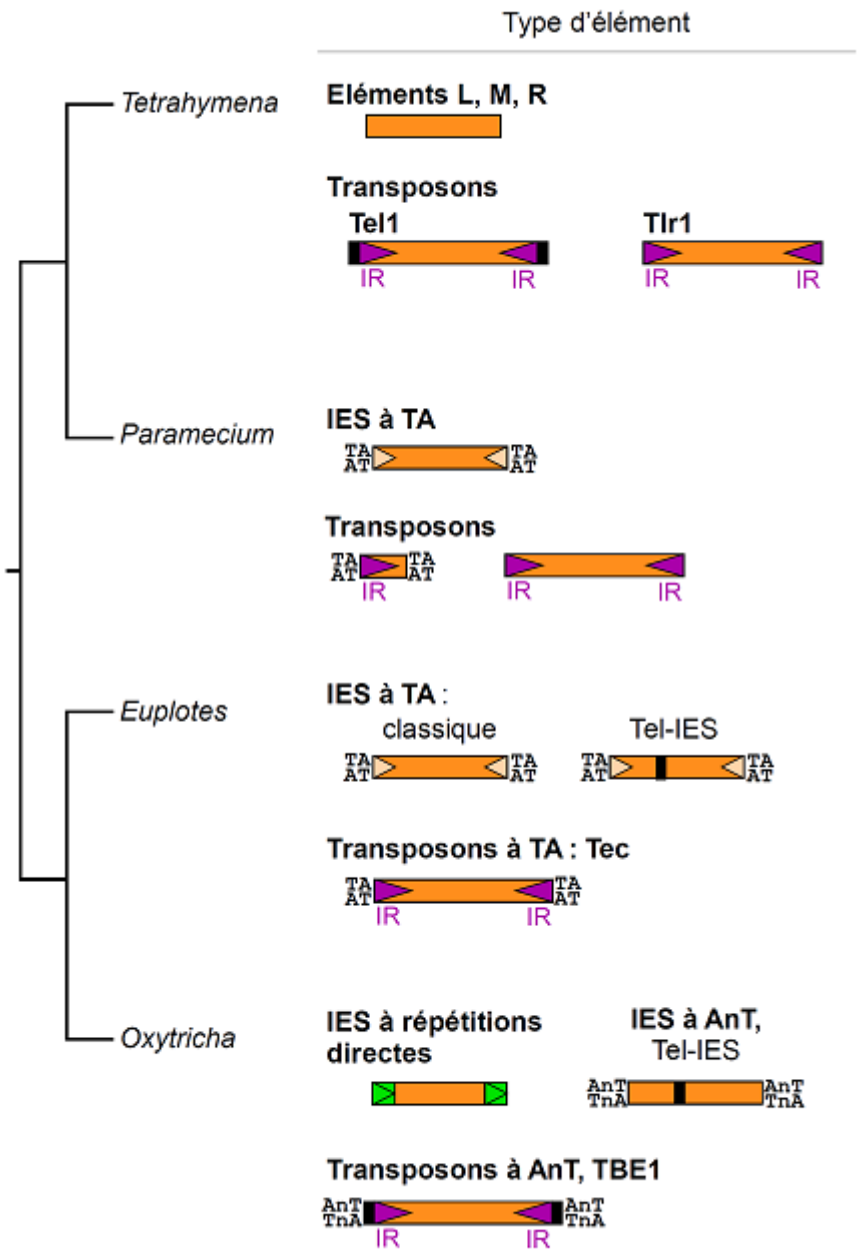


Figure 28 : Résumé des différents éléments éliminés pendant les réarrangements programmés des génomes de *Paramecium, Tetrahymena, Euplotes* et *Oxytricha.*

La situation chez les autres ciliés.

Les réarrangements sont présents chez tous les ciliés, cependant ils présentent des différences selon les organismes. La nature des éléments éliminés (figure 28), et leur position dans le génome varient d'un organisme à l'autre. Si *Tetrahymena thermophila* possède des mécanismes relativement proches de ceux présents chez *Paramecium tetraurelia*, d'autres comme *Oxytricha* voient leur génome subir des réarrangements extrêmes.

Tetrahymena

Tetrahymena thermophila possède un génome MAC d'environ 100Mpb. Ses génomes MAC et mic ont été séquencés et annotés. Le génome mic n'est pas encore disponible. Cependant quelques milliers d'éléments spécifiques de la lignée germinale ont été extrapolés à partir de l'étude de quelques régions micronucléaires. Parmi ces éléments d'une taille pouvant aller de 600 pb à 21kb, certains sont appelés IES bien qu'ils ne partagent pas l'ensemble des caractéristiques propres aux IES de *Paramecium tetraurelia.* C'est le cas des éléments L, M et R. Bien qu'appelés IES dans la littérature, ces éléments se rapprochent plus des séquences éliminées de façon imprécise chez *Paramecium*. Ils sont non codants, répétés le long du génome, plus riche en AT que le reste du génome et n'interrompent pas les cadres ouverts de lecture. On n'observe pas de séquence unique au niveau des jonctions, mais des séquences hétérogènes car des séquences alternatives peuvent être employées comme bornes pour l'introduction des cassures double brin programmées (Saveliev and Cox, 2001).

D'autres séquences sont éliminées chez *Tetrahymena,* il s'agit des transposons Tel-1 et Tlr. Ils possèdent des répétitions inversées répétées qui ne servent pas de bornes de délétion. On peut trouver dans les éléments Tlr des cadres ouvert de lecture de protéines de transposons. Les six éléments Tel-1 font entre 9 et 13kb tandis que les 30 éléments Tlr ont une taille comprise entre 13 et 21kb.

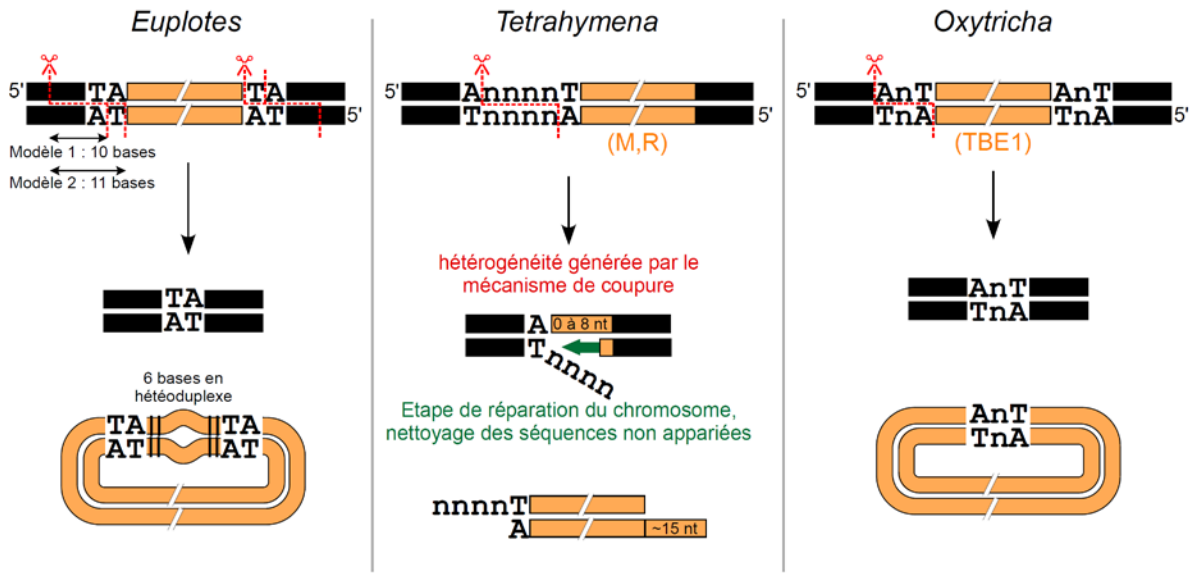


Figure 29 : Différents produits de réparation au niveau des jonctions chromosomiques et des éventuels cercles d'IES chez Euplotes (Jahn and Klobutcher, 2002), Tetrahymena (Saveliev and Cox, 2001) et Oxytricha (Williams et al., 1993). Les éléments éliminés apparaissent en orange et l'ADN flanquant en noir.

Une dizaine d' IES sont flanquées par des répétitions TTAA, interrompent potentiellement des exons et sont précisément excisées (Fass et al., 2011).

Une transposase domestiquée de type *piggybac* a été identifiée chez *Tetrahymena*. Appelée Tpb2, cette protéine est exprimée pendant les réarrangements. Lorsqu'elle est déplétée par ARN interférence on observe une inhibition des éliminations d'ADN. Des études in vitro ont montré qu'il s'agissait d'une endonucléase capable de produire des cassures double brin avec une géométrie similaire à celle observée in vivo chez *Tetrahymena* (Saveliev and Cox, 2001), quatre bases sortantes en 5' sans qu'un TA central soit requis, de plus les jonctions d'excision sont hétérogènes (Cheng et al., 2010).

Euplotes

Euplotes crassus élimine pendant les réarrangements 90% de son génome. *Euplotes* possède des IES de TA similaires à celles présente chez *Paramecium*. Elles sont courtes, typiquement entre 30 et 400 pb, elles possèdent des dinucléotides TA à leurs bornes, peuvent interrompre des cadres ouverts de lecture et sont excisées de façon précise.

Euplotes excise également plusieurs dizaines de milliers de séquences plus proches des transposons, les éléments Tec. Ils sont plus longs que les IES, en effet ils ont une taille d'environ 5kb. De plus ils possèdent des répétitions inversées terminales d'environ 700 pb et sont encadrés par de courtes répétitions directes d'un dinucléotide TA. Une grande partie d'entre eux peuvent également interrompre des cadres ouverts de lecture. Les Tec sont éliminés en deux vagues. Pendant la première vague, l'élimination est précise et ressemble à l'excision des IES. Lors de la deuxième vague, l'élimination est imprécise et peut mener à la fragmentation.

L'excision des IES et l'élimination précise des Tec (lors de la première vague) sont très similaires. En effet, toutes deux sont très précises et génèrent des cercles d'ADN excisés. De plus au niveau de la jonction macronucléaire, une seule copie de la répétition directe TA est

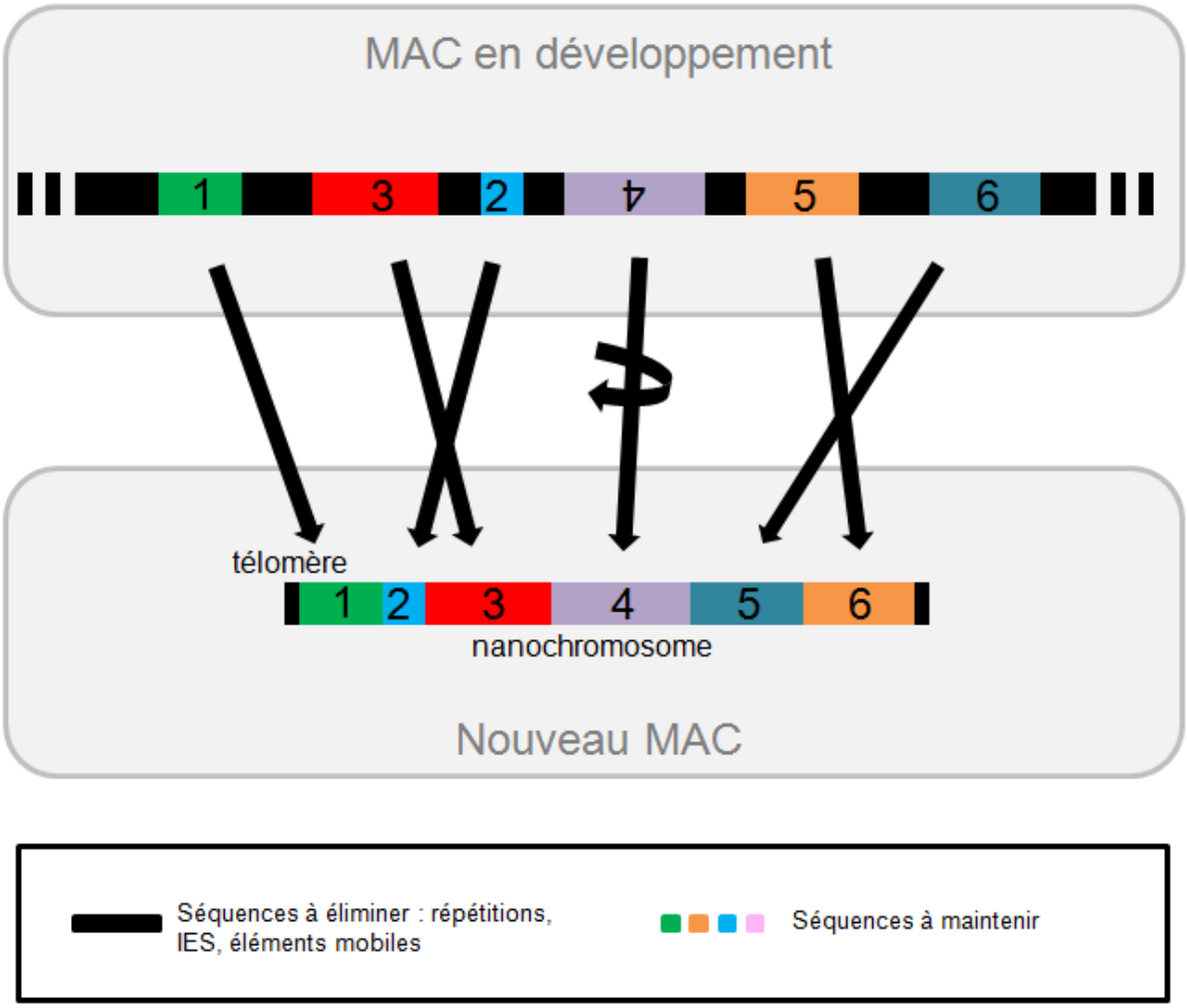


Figure 30 : Unscrambling chez *Oxytricha*. Les fragments géniques à maintenir doivent etre remis dans le bon ordre et le bon sens pour produire un nanochromosome portant un gène fonctionnel.

maintenue. Cela suggère que ces éléments partagent un mécanisme d'élimination commun ou deux systèmes d'éliminations distincts qui ont des protéines communes.

Une observation surprenante est la différence de structure observé au niveau de la réparation des jonctions macronucléaires et des cercles d'ADN excisé. Si la jonction macronucléaire est extrêmement précise, la jonction des cercles est particulière, elle consiste en deux répétitions directes contenant le dinucléotide TA encadrant 10 pb comportant 6 nucléotides en hétéroduplex. Les deux mécanismes de réparation au niveau de ces cassures sont donc différents. L'analyse des produits de jonction d'excision ont conduit à imaginer un modèle dans lequel les cassures génèrent une extrémité de 10 bases sortantes en 5'. Coté jonction macronucléaire, des extrémités s'apparieraient au niveau du dinucléotide TA puis les gaps seraient comblés par une étape de polymérisation avant une religation des extrémités par une ligase ADN. Coté cercle d'ADN excisé, on imagine que les extrémités seraient prises en charge par le complexe d'excision lui-même ou par un complexe de réparation puis rapprochées sans appariement, polymérisation pour combler les gaps puis ligation. Ce qui génère un cercle avec un 10 pb encadrées par les dinucléotides TA, contenant un hétéroduplex de 6 pb (Jaraczewski and Jahn, 1993; Klobutcher et al., 1993). La nucléase responsable des réarrangements n'a pas été identifié.

Oxytricha

Ce cilié apparenté à *Euplote*s dévoile des mécanismes extrêmement complexes pendant les réarrangements de son génome. Non seulement les IES doivent être excisées de façon précise mais les séquences qui formeront le nouveau MAC (MDS) doivent être remises dans le bon ordre. Dans cet organisme les fragments de gènes séparés par les IES, sont mélangés, dans un ordre incorrect, peuvent être dans la mauvaise orientation et parfois dans certains cas extrêmes, se trouver sur un autre chromosome (figure 30). De plus les réarrangements produisent des nanochromosomes de la taille d'un seul gène avec des télomères au niveau

des extrémités. Pendant le processus, 90 à 95% de l'ADN germinal va être éliminé.

Les séquences spécifiques de la lignée germinale chez *Oxytricha trifallax* sont de plusieurs types.

Les éléments TBE, au nombre de 2000, ressemblent à des transposons, contiennent des cadres ouverts de lecture codant pour des transposases, ils possèdent des répétitions inversées d'environ 70 pb, et présentent des homologies avec des transposons Tc1/mariner. Cependant les répétitions directes à leurs bornes diffèrent. Ce ne sont pas des dinucléotides TA, mais des répétitions directes ANT. Les jonctions macronucléaires sont homogènes, l'excision est précise et laisse une répétition ANT dans le nouveau génome MAC

D'autres éléments peuvent être assimilés à des IES. Ces séquences sont courtes et ne possèdent pas de répétitions inversées caractéristiques des transposons et ne codent pas pour une transposase. Elles sont uniques, riches en AT et la grande majorité d'entre elles ont une taille inférieure à 100 pb. Elles sont excisées de façon précise bien qu'elles n'aient pas toutes les mêmes bornes. Elles peuvent contenir des dinucléotides TA mais ce n'est pas le cas de la totalité d'entre elles.

Plusieurs transposases à domaine DDE ont été trouvées dans les génomes d'Oxytricha. Des transposases de la famille Tc1/mariner codées par les transposons TBE sont exprimées pendant la différentiation du macronoyau d'*Oxytricha* et sont nécessaire pour les réarrangements de l'ADN (Nowacki et al., 2009).

D'autres gènes portés par le macronoyau, expriment des protéines qui pourraient être des transposases domestiquées, sans qu'il soit encore déterminé si elles sont impliquées dans les réarrangements. Ces gènes contiennent soit un domaine MULE, dérivé de transposons Mutator-like, dans 11 cas distincts, soit un domaine DDE_Tnp_IS1595 dans 9 cas. Ils sont exprimés pendant la conjugaison mais plus tardivement que la transposase encodée par les

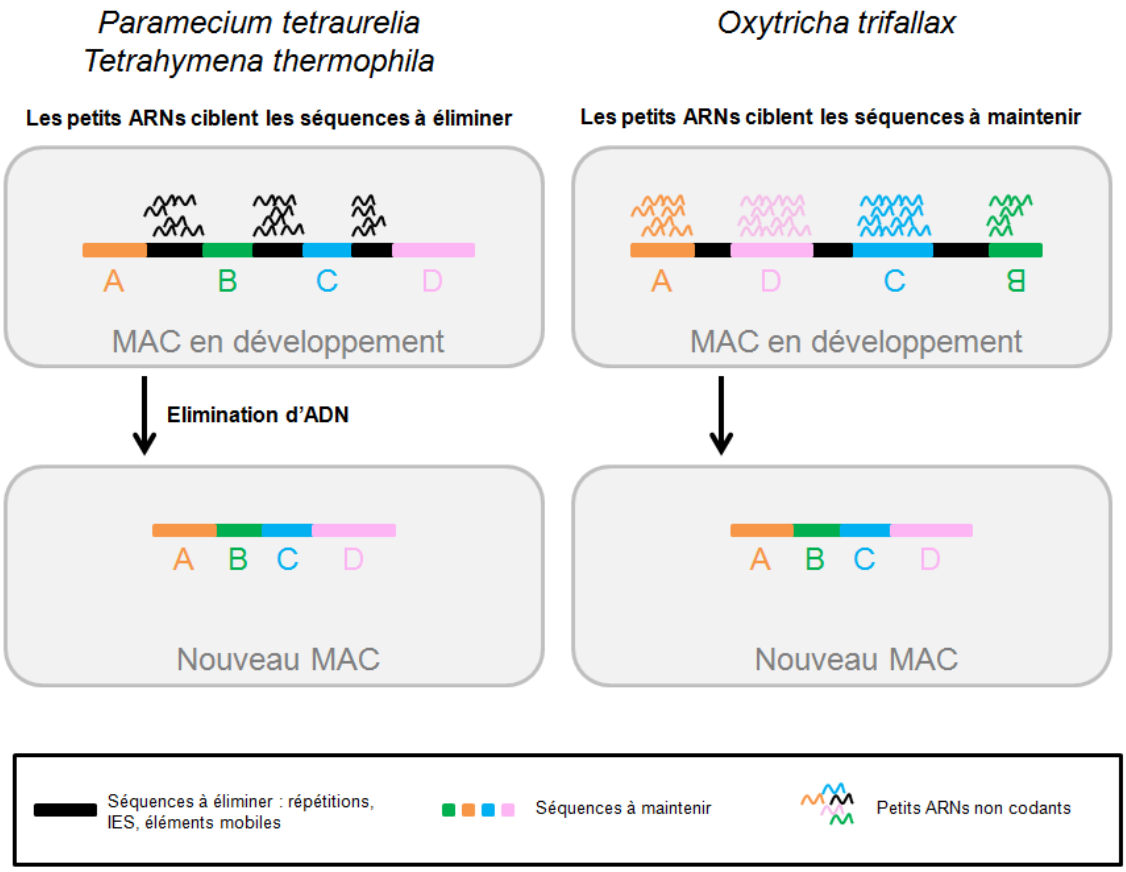


Figure 31 : Ciblage des séquences à maintenir. Chez *Oxytricha*, les petits ARNs ciblent les séquences à maintenir et le « unscrambling » rétablit le bon ordre dans les séquences géniques pour produire un gène fonctionnel. (Sontheimer, 2012)

éléments TBE. Il est important de noter que tous ces gènes ne possèdent pas forcément un site catalytique DDE intègre (Swart et al., 2013; Williams et al., 1993).

Il est encore difficile de savoir quelles transposases sont impliquées dans les réarrangements et les signaux permettant de les recruter. Cependant il est possible d'imaginer la collaboration de plusieurs mécanismes, la transposase des transposons TBE excisant les transposons, tandis qu'une ou des transposases domestiquées MULE ou IS1595 pourraient être nécessaires pour l'élimination des éléments proches des IES de *Paramecium tetraurelia.* De façon surprenante, le mécanisme de contrôle des réarrangements par les ARNs non codants semble inversé chez *Oxytricha*. Les petits ARNs scan ciblent les séquences à conserver, qui sont très minoritaires dans le génome micronucléaire (figure 31) (Fang et al., 2012).

Projet

Au début de ce travail de thèse, les travaux d'Aurélie Kapusta sur le rôle de la Ligase IV et de son partenaire XRCC4 touchaient à leur terme. Il avait été montré que ces deux protéines sont indispensables pour une progression normale des réarrangements programmés des génomes. Plus particulièrement, lorsque ces protéines sont absentes, les cassures double brin programmées sont introduites normalement aux bornes des IES, avec le même timing qu'en conditions contrôles mais ne sont pas réparées et s'accumulent. Cependant, ces extrémités d'ADN cassées ne semblent pas dégradées. L'excision des IES impliquant un acteur majeur de la voie NHEJ, nous avons proposé que l'hétérodimère Ku70/80 était présent au niveau des extrémités cassées et les protégeait de la dégradation.

Il a été observé que la protéine Pgm est nécessaire, de façon directe ou indirecte, pour tous les types de réarrangements. Cependant les mécanismes discriminant entre réarrangements précis et imprécis restaient inconnus.

L'élimination hétérogène d'éléments répétés, caractérisée par des jonctions de réparation imprécise, la réparation après élimination des éléments répétés pourrait être prise en charge par une voie de recollement des extrémités cassées indépendante de l'hétérodimère Ku70/80, la voie alt-NHEJ

Dans ce mémoire je présente les travaux réalisés lors de ma thèse. Tous concernent les réarrangements programmés du génome de *Paramecium tetraurelia* et peuvent être organisés en trois parties :

- L'identification par séquençage haut débit, et la validation par biologie moléculaire d'un jeu de référence de 45 000 IES à partir d'ADN de paramécie dont le gène PGM a été éteint. A cela s'ajoute une analyse des séquences retrouvées aux bornes des IES.

- L’étude du rôle de la ligase IV et son partenaire XRCC4 lors de l’excision des IES et leur influence sur la maturation des extrémités cassées. Dans la recherche des acteurs de la maturation des extrémités, je me suis intéressé également aux protéines CtIP et Mre11.

- L’analyse des gènes KU de la paramécie, l’effet de leur extinction sur la survie de la descendance, et la description les mécanismes moléculaires lors de l’excision des IES. Les résultats pour la caractérisation des réarrangements imprécis sont également présentés ici.

Enfin je discuterai de l’implication des résultats obtenus sur l’excision des IES et de quelle manière ils changent notre vision du modèle de réarrangements programmés du génome chez *Paramecium tetraurelia* et de l’étroite relation entre réarrangements programmés et réparation.

RESULTATS

PAPIER IES + résultats supplémentaires

Les IES de paramécie dérivent de transposons

Pendant longtemps, le génome micronucléaire n'était pas séquencé, les micronoyaux représentant une fraction trop minoritaire de l'ADN contenu par les paramécies pour être séquencé. L'étude des IES et de leur excision était menée sur un jeu de quelques dizaines d'IES. A partir des IES connues, une densité d'IES a été extrapolée, et bien qu'à partir de peu de séquences, une grande similarité a été observée entre un consensus des bornes des IES de *Paramecium* et celui des éléments Tc1/mariner. L'hypothèse a été émise que les IES sont des traces d'anciens transposons, et le modèle IBAF a été proposé.

Séquençage du MAC de cellules déplétées pour Pgm

Il a été montré que la transposase domestiquée Pgm était responsable pour l'introduction des cassures double brin programmées aux bornes de toutes les IES testées. En son absence l'élimination des quelques transposons connus était inhibée également bien qu'on ne sache pas s'il s'agissait d'un effet direct ou indirect. Cependant, même si les réarrangements sont impossibles, l'extinction de cette protéine n'empêche pas l'amplification de l'ADN. Cela a laissé entrevoir la possibilité d'accéder à la séquence des micronoyaux par le séquençage des macronoyaux de cellules déplétées pour Pgm.

Les IES dérivent de transposons

Le séquençage a effectivement permis de séquencer et d'identifier un jeu de 45000 IES de référence. A partir de ces données il a été possible de confirmer les tendances qui jusqu'ici n'étaient basées que sur moins d'une centaine d'IES. Cela nous a permis également de tester l'hypothèse selon laquelle les IES seraient des traces d'anciens transposons ayant envahi le génome de la paramécie.

Ce travail présenté ici a été réalisé avec de nombreux laboratoires travaillant sur la paramécie. Mon travail a principalement consisté en la validation par biologie moléculaire des résultats obtenus *in silico.*

OPEN ACCESS Freely available online

PLOS | GENETICS

The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences

Olivier Arnaiz[1,2,3], Nathalie Mathy[1,2,3], Céline Baudry[1,2,3], Sophie Malinsky[4,5,6], Jean-Marc Aury[7], Cyril Denby Wilkes[1,2,3], Olivier Garnier[4,5,6], Karine Labadie[7], Benjamin E. Lauderdale[8], Anne Le Mouël[4,5,6¤], Antoine Marmignon[1,2,3], Mariusz Nowacki[9], Julie Poulain[7], Malgorzata Prajer[10], Patrick Wincker[7,11,12], Eric Meyer[4,5,6], Sandra Duharcourt[13], Laurent Duret[14], Mireille Bétermier[1,2,3]*, Linda Sperling[1,2,3]*

1 CNRS UPR3404 Centre de Génétique Moléculaire, Gif-sur-Yvette, France, **2** Département de Biologie, Université Paris-Sud, Orsay, France, **3** CNRS FRC3115, Centre de Recherches de Gif–sur-Yvette, Gif-sur-Yvette, France, **4** Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, France, **5** INSERM, U1024, Paris, France, **6** CNRS, UMR 8197, Paris, France, **7** Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, Evry, France, **8** Methodology Institute, London School of Economics, London, United Kingdom, **9** Institute of Cell Biology, University of Bern, Bern, Switzerland, **10** Department of Experimental Zoology, Institute of Systematics and Evolution of Animals, Polish Academy of Sciences, Krakow, Poland, **11** Centre National de Recherche Scientifique (CNRS), UMR 8030, CP5706, Evry, France, **12** Université d'Evry, Evry, France, **13** Institut Jacques Monod, CNRS, UMR 7592, Université Paris Diderot, Sorbonne Paris Cité, Paris, France, **14** Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France

Abstract

Insertions of parasitic DNA within coding sequences are usually deleterious and are generally counter-selected during evolution. Thanks to nuclear dimorphism, ciliates provide unique models to study the fate of such insertions. Their germline genome undergoes extensive rearrangements during development of a new somatic macronucleus from the germline micronucleus following sexual events. In *Paramecium*, these rearrangements include precise excision of unique-copy Internal Eliminated Sequences (IES) from the somatic DNA, requiring the activity of a domesticated *piggyBac* transposase, PiggyMac. We have sequenced *Paramecium tetraurelia* germline DNA, establishing a genome-wide catalogue of ~45,000 IESs, in order to gain insight into their evolutionary origin and excision mechanism. We obtained direct evidence that PiggyMac is required for excision of all IESs. Homology with known *P. tetraurelia* Tc1/mariner transposons, described here, indicates that at least a fraction of IESs derive from these elements. Most IES insertions occurred before a recent whole-genome duplication that preceded diversification of the *P. aurelia* species complex, but IES invasion of the *Paramecium* genome appears to be an ongoing process. Once inserted, IESs decay rapidly by accumulation of deletions and point substitutions. Over 90% of the IESs are shorter than 150 bp and present a remarkable size distribution with a ~10 bp periodicity, corresponding to the helical repeat of double-stranded DNA and suggesting DNA loop formation during assembly of a transpososome-like excision complex. IESs are equally frequent within and between coding sequences; however, excision is not 100% efficient and there is selective pressure against IES insertions, in particular within highly expressed genes. We discuss the possibility that ancient domestication of a *piggyBac* transposase favored subsequent propagation of transposons throughout the germline by allowing insertions in coding sequences, a fraction of the genome in which parasitic DNA is not usually tolerated.

Citation: Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury J-M, et al. (2012) The *Paramecium* Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences. PLoS Genet 8(10): e1002984. doi:10.1371/journal.pgen.1002984

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received March 16, 2012; **Accepted** August 9, 2012; **Published** October 4, 2012

Funding: This work was supported by the ANR BLAN08-3_310945 "ParaDice," the ANR 2010 BLAN 1603 "GENOMAC," a CNRS ATIP-Plus grant to MB (2010–2011), and an "Equipe FRM" grant to EM. The sequencing was carried out at the Genoscope - Centre National de Séquençage (Convention GENOSCOPE-CEA number 128/AP 2007_2008/CNRS number 028666). CDW and AM were supported by Ph.D. fellowships from the Ministère de l'Enseignement Supérieur et de la Recherche. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mireille.betermier@cgm.cnrs-gif.fr (MB); linda.sperling@cgm.cnrs-gif.fr (LS)

¤ Current address: UMR7216 Epigénétique et Destin Cellulaire, CNRS, Université Paris-Diderot/Paris 7, Paris, France

Introduction

Paramecium belongs to the ciliate phylum, a deep radiation of highly diverse unicellular eukaryotes. The hallmark of ciliates is nuclear dimorphism: each unicellular organism harbors two kinds of nuclei with distinct organization and function. A diploid "germline" micronucleus (MIC) undergoes meiosis and transmits the genetic information to the next sexual generation but is not expressed. A polyploid "somatic" macronucleus (MAC) contains a version of the genome streamlined for gene expression and determines the phenotype. A new MAC is formed at each sexual generation by programmed rearrangements of the entire zygotic, germline-derived genome, and the maternal MAC is lost. The MAC genome of *P. tetraurelia* has been sequenced [1] revealing a

Author Summary

Ciliates are unicellular eukaryotes that rearrange their genomes at every sexual generation when a new somatic macronucleus, responsible for gene expression, develops from a copy of the germline micronucleus. In *Paramecium*, assembly of a functional somatic genome requires precise excision of interstitial DNA segments, the Internal Eliminated Sequences (IES), involving a domesticated *piggyBac* transposase, PiggyMac. To study IES origin and evolution, we sequenced germline DNA and identified 45,000 IESs. We found that at least some of these unique-copy elements are decayed Tc1/mariner transposons and that IES insertion is likely an ongoing process. After insertion, elements decay rapidly by accumulation of deletions and substitutions. The 93% of IESs shorter than 150 bp display a remarkable size distribution with a periodicity of 10 bp, the helical repeat of double-stranded DNA, consistent with the idea that evolution has only retained IESs that can form a double-stranded DNA loop during assembly of an excision complex. We propose that the ancient domestication of a *piggyBac* transposase, which provided a precise excision mechanism, enabled transposons to subsequently invade *Paramecium* coding sequences, a fraction of the genome that does not usually tolerate parasitic DNA.

series of whole genome duplications (WGDs) in the lineage that provide a unique tool for evolutionary analyses.

Ciliate genome rearrangements and their epigenetic control by non-coding RNAs have been recently reviewed [2–4]. In *Paramecium*, genome rearrangements involve (i) endoreplication of the DNA to about 800 haploid copies, (ii) imprecise elimination of genomic regions that contain, in particular, transposons and other repeated sequences, usually leading to chromosome fragmentation and (iii) elimination of Internal Eliminated Sequences (IES) by a precise mechanism. The accuracy of this process is crucial for IESs located within coding regions, to correctly restore open reading frames. The characterization of fewer than 50 IESs identified by cloning MIC loci [5] showed that they are short (26–883 bp), unique copy elements that are located in both coding and non-coding regions of the genome. The IESs are invariably flanked by two TA dinucleotides whereas only one TA is found at the MAC chromosome junction after IES excision (Figure 1). IESs have also been discovered by *cis*-acting mendelian mutations that prevent their excision, conferring a mutant phenotype [6–10]. The mutations in almost all cases were found in one of the flanking TA dinucleotides, which seem to be an absolute sequence requirement for IES excision. Extrapolation of the number of IESs found mainly in surface antigen genes led to the estimation that there could be as many as 50,000 IESs in the *Paramecium* genome. Such massive presence of unique copy IESs inserted in genes is not a characteristic of all ciliates. The estimated 6,000 IESs of the related oligohymenophorean ciliate *Tetrahymena* [11] are excised by an imprecise mechanism [12], are usually multi-copy including recognizable transposons [13–15] and are rarely found in coding sequences [16,17].

Klobutcher and Herrick [18] first reported a weak consensus at the ends of 20 IESs from *Paramecium* surface antigen genes (5′-TAYAGYNR-3′) that resembles the extremities of Tc1/mariner transposons. These authors hypothesized a "transposon link" to explain the origin of IESs, suggesting that they are the decayed relics of a Tc1/mariner transposon invasion and that they are excised from the MAC DNA by a Tc1/mariner transposase encoded by a gene that has become part of the cellular genome [19]. In this model, IES excision represents the exact reversal of Tc1/mariner transposon integration into its TA target site with duplication of the TA dinucleotide, an evolutionary novelty that may have appeared more than once in the ciliate phylum. One problem with the model is that transposition catalyzed by Tc1/mariner transposases usually leaves a 2 or 3 bp "footprint" at the donor site [20] while IES excision is precise.

A decisive step towards understanding the mechanism of IES excision and validating a transposon link for the origin of the IES excision machinery was the identification of a domesticated *piggyBac* transposase in *Paramecium* [21]. Baptized PiggyMac (Pgm), the protein is encoded by the *PGM* gene which is expressed only late in sexual processes, at the time of genome rearrangements. Pgm, localized in the developing new MAC, was found to be required for the excision of all IESs tested and for the imprecise elimination of several regions containing transposons or cellular genes [21]. A similar *piggyBac*-derived transposase is found in *Tetrahymena* and is required for heterochromatin-dependent DNA elimination [22]. Since the *Paramecium* and *Tetrahymena* proteins appear to be monophyletic, based on a broad phylogeny of *piggyBac* transposases (L. Katz and F. Gao, personal communication), the domestication event may have preceded the divergence of these two ciliates, estimated at 500–700 Ma (million years ago) [23]. Most significantly, the *in vivo* geometry of IES excision, initiated by staggered double-strand breaks (DSBs) that generate 4-base 5′ overhangs centered on the TA at both ends of the IES

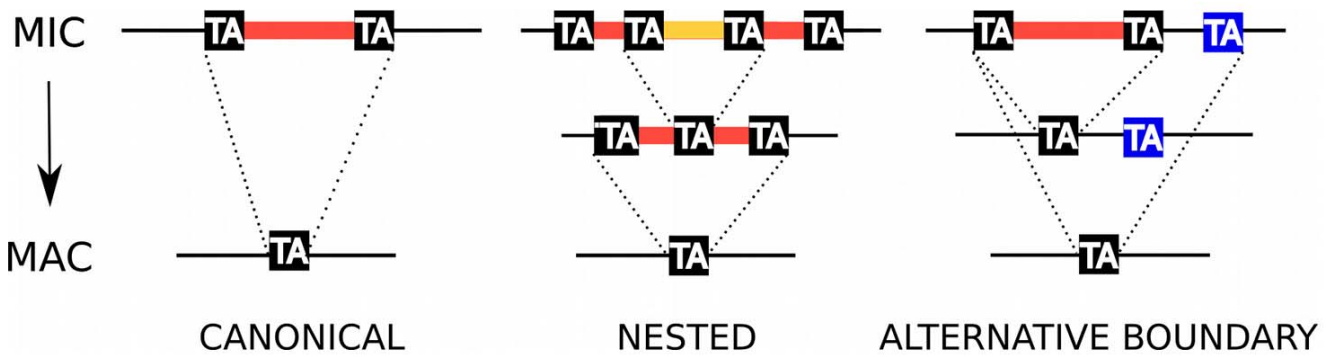


Figure 1. IES excision. Schematic representation of, from left to right, a canonical IES, a nested IES and an IES with an alternative boundary. In the case of the nested IES, the middle line represents either an intermediate in the excision pathway or an alternative final product. In the case of the alternative boundary IES, the middle line represents an alternative final product.
doi:10.1371/journal.pgen.1002984.g001

[24], is fully compatible with the *in vitro* reaction catalyzed by a *piggyBac* transposase isolated from an insect [25], whose target site is a 5′-TTAA-3′ tetranucleotide. *piggyBac* elements leave behind no scar when they jump to a new location: only ligation is required to join the fully complementary 5′ overhangs. Limited processing of 5′ and 3′ ends is further required for precise closure of the *Paramecium* IES excision sites since only the TA dinucleotides at the center of the 4-base 5′ overhangs are always complementary [24,26].

We report here a genomic approach to exhaustively catalogue the IESs in the *Paramecium tetraurelia* germline genome in order to study their evolutionary dynamics and seek evidence for a transposon origin of these elements. We obtained DNA highly enriched in un-rearranged germline sequences, from cells depleted in Pgm by RNA interference. Deep-sequencing of this DNA (hereafter called "PGM DNA") allowed us to identify a genome-wide set of nearly 45,000 IESs, by comparing contigs assembled using the PGM DNA (hereafter called "PGM contigs") with the MAC reference genome [1]. The hypothesis that Pgm is required for excision of all IESs was tested by genome-scale sequencing of a source of DNA from purified MICs [27], providing validation of the IES catalogue. The evolutionary dynamics of the IESs was studied by exploiting the series of WGDs that have been characterized in *Paramecium* [1]. The study provides, to our knowledge, the first genome-wide set of IESs, in *Paramecium* or any ciliate, and provides new evidence that IESs have deleterious effects on fitness and that at least a fraction of IESs do derive from Tc1/mariner transposons that have decayed over time. The IES sequences evolve rapidly. The constraints we could detect concern their size distribution, suggestive of the assembly of a transpososome-like excision complex and a weak consensus at their ends, which resembles the extremities of Tc1/mariner elements. We discuss the possibility that ancient domestication of the Pgm transposase favored subsequent propagation of transposons throughout the *Paramecium* germline genome, by providing a mechanism for their precise somatic excision, therefore allowing insertions in coding sequences.

Results

IES identification

An overview of the strategy for identification of a genome-wide set of IESs is presented in Figure S1. The first step was next-generation deep sequencing of DNA enriched in un-rearranged sequences, isolated from strain 51 cells that had undergone the sexual process of autogamy after depletion of Pgm protein by RNAi (Figure S2). In the absence of Pgm, the zygotic DNA is amplified but rearrangements are impaired. The sample that was sequenced contained a mixture of 60–65% un-rearranged DNA from the developing new MACs and 35–40% rearranged DNA from the fragments of the maternal MAC still present in the cytoplasm, as judged by Southern blot quantification of MIC and MAC forms at one locus (Figure S3). The PGM sequence reads (Table 1) were mapped to the MAC reference genome of strain 51 (see Materials and Methods), and putative IES insertion sites were defined as sites with a local excess of ends of read alignments (pipeline MIRAA for "Method of Identification by Read Alignment Anomalies"). This excess of ends of alignments arises when a read contains a MIC IES junction, since only part of such a read can align with a MAC chromosome, either starting or ending at the IES insertion site, expected to be a TA dinucleotide. Using MIRAA, we identified 45,739 potential IES insertion sites. Essentially all (99%) of the insertion sites contained a TA dinucleotide, even though this was not assumed by the pipeline.

In order to obtain the sequence of the IESs, the paired-end PGM DNA sequence reads were assembled into contigs (cf. Table S1 for assembly statistics) and compared to the MAC reference genome assembly (pipeline MICA for "Method of Identification by Comparison of Assemblies"). We looked for insertions in the PGM contigs with respect to the MAC reference assembly. Any insertion bounded by TA dinucleotides after local realignment was considered to be an IES. Using this pipeline we identified 44,928 IESs. The fact that 96% (n = 43,220) of the IESs identified by MICA correspond to an IES insertion site identified by MIRAA (Figure S1) testifies to the overall reliability of the procedure. Experimental validation of 6 IESs identified only by MICA and 17 insertion sites identified only by MIRAA was carried out by PCR amplification of an independent preparation of PGM DNA. The results (Table S2) show that the 6 IESs and at least 12 of 17 insertion sites tested do correspond to the presence of an IES. Interestingly, among the IES sites identified only by MIRAA, we found 8 examples of a pair of IESs separated by one or only a few nucleotides (in 5/8 cases, these tandem IESs are located in exons, a proportion similar to that found for the genome-wide IES set, see below). This case is not handled by the MICA pipeline since the initial global alignment with BLAT would have detected a single large insertion that would have been rejected by the local realignment filter, which requires the insertion to be flanked by TA dinucleotides. This is the first report of such closely spaced IESs, although nested IESs (Figure 1) have been previously documented [8].

In order to see whether the set of 44,928 IESs is likely to be exhaustive, we looked for the 53 previously characterized IESs identified directly by cloning MIC loci in *P. tetraurelia* strain 51 cells (Table S3). All 53 previously cloned IESs were found, with the exception of one IES that had been assembled into the MAC reference genome and one IES form that represents use of an alternative boundary. In addition, two small IESs, each of which is

Table 1. Sequencing and mapping statistics.

DNA	Insert size (bp)	Read length (bp)	Reads	Aligned reads	Aligned (%)	Genome coverage (%)
PGM	~500	108	130,266,728	110,189,736	84.6	99
Lambda-phage	~200	101	83,149,385	25,949,607	31	44

Paired-end Illumina sequencing was carried out as described in Materials and Methods, and reads were mapped to the *P. tetraurelia* MAC reference genome using the BWA short-read aligner. The genome coverage is the fraction of the genome covered by at least 1 read. The depth of coverage with the PGM DNA is on average 165×. The depth of coverage with the lambda-phage DNA is on average 75× for the part of the genome that is covered. The PGM reads that were not aligned contain *Paramecium* mitochondrial and rDNA sequences, contaminating bacterial sequences as well as sequences present only in the MIC genome. In addition, a large proportion of the unaligned lambda-phage reads are from bacterial contaminants with AT-rich genomes; this DNA was not eliminated by the cesium chloride density gradient separation step of the phage library construction [28].

doi:10.1371/journal.pgen.1002984.t001

nested within a larger IES, were found in PGM DNA but were not identified by our pipeline as IESs. Indeed, nested IESs can only be identified by time-course experiments or if the outer IES is retained in the MAC e.g. as the result of a point mutation [8]. Since 49 of 51 non-nested IESs were identified by MICA, the IES identification procedure has a sensitivity of at least 96%.

The entire IES identification approach is based on the assumption that the excision of all IESs in *Paramecium* requires the Pgm domesticated transposase activity. In order to test this assumption, we sequenced inserts from a lambda-phage library constructed some 20 years ago [28], using DNA from MICs that had been separated from MACs by Percoll gradient centrifugation [27]. This library has been extensively used to clone MIC loci with specific probes. Although the contigs assembled from the phage DNA reads only partially covered the MAC reference genome (Table 1), 98.5% of the 13,377 IESs that could be identified using the phage DNA and the MICA pipeline had also been identified using the PGM DNA. The difference of 1.5% is within the estimated sensitivity of the MICA pipeline. We conclude that all *Paramecium* IESs very likely require Pgm for excision, and that our data set does represent a genome-wide set of *P. tetraurelia* IESs.

IES distribution in the genome

The genome-wide set of IESs has an overall G+C content of 20%, significantly lower than the 28% G+C content of the MAC reference genome [29] but comparable to the G+C content of intergenic regions (21%). The IESs are found in exons (76.8%), introns (5.4%) and intergenic regions (17.8%), suggesting a nearly random distribution of IESs with respect to genes, since the MAC reference genome is composed of 76% exons, 3.2% introns and 20.8% intergenic DNA [1]. However, IESs are not randomly distributed along the chromosomes. Intriguingly, as shown in Figure S4 for the 8 largest MAC chromosomes, IESs tend to be asymmetrically distributed along MAC chromosomes. The MAC assembly (188 scaffolds >45 Kb constitute 96% of the 72 Mb assembly) contains 115 telomere-capped scaffolds, varying in size from ~150 Kb to ~1 Mb, that are considered to represent complete MAC chromosomes. For 70 of these telomere-capped scaffolds, IESs display non-uniform distributions ($p<0.002$, median scaffold size 417 Kb) while for the remaining 45 telomere-capped scaffolds, the IES distribution is uniform (median scaffold size 275 Kb). Thus the larger the MAC chromosome, the greater the chance of observing a non-uniform IES distribution. The distributions for all scaffolds are easily visualized using the ParameciumDB [30] Genome Browser. The significance of the asymmetry in IES distribution is not clear, but might be related to the global organization of MIC chromosomes, currently unknown (discussed in [29]).

Germline Tc1/mariner transposons

The genome-wide set of IESs covers 3.55 Mb (mean IES size 79 bp), compared to 72 Mb for the MAC reference genome assembly. The IESs thus add about 5% to the sequence complexity of the part of the MIC genome that is collinear with MAC chromosomes. The total complexity of the PGM contigs (after elimination of contigs with low PGM read coverage and high G+C content, assumed to represent bacterial contamination as confirmed in many cases by BLASTN matches against bacterial genomes) is ~100 Mb, however the use of a single paired-end sequencing library with small inserts (~500 bp) may have perturbed assembly of repeated sequences, possibly leading to underestimation of repeated sequence content. We infer that ~25 Mb of germline-specific DNA corresponds to the imprecisely eliminated regions located outside of the MAC-destined chromosomes i.e. the part of the MIC genome that is not collinear with MAC chromosomes.

We have not further characterized this fraction of the PGM DNA. However, we did identify the first germline *P. tetraurelia* Tc1/mariner transposons (Figure S5), by using the phage-lambda library of MIC DNA [28] to walk past the end of MAC scaffold_51, which bears the subtelomeric 51G surface antigen gene [31]. In all, 5 phage inserts and 4 cloned PCR products corresponding to part or all of different copies of the element downstream of the 51G surface antigen gene, named *Sardine*, were sequenced (EMBL Nucleotide Sequence Database accession numbers HE774468–HE774475) and a consensus for the ~6.7 Kb transposon was constructed (Figure S5 and Text S1). The ends of the *Sardine* copies contain intact or partially deleted 425 bp terminal inverted repeats (TIRs) which are themselves palindromic, containing a unique, oriented region nested within outer inverted repeats (Figure S5). *Sardine* contains up to 4 ORFs. One ORF is a putative DD35E transposase of the IS*630*-Tc1 family, like the DDE transposases of the TBE and Tec transposons found in stichotrich ciliates [32]. Another ORF, as in Tec transposons [33], encodes a putative tyrosine recombinase. The other two ORFs are hypothetical, though ORF2 shows some similarity (31.7% identity and 55.4% similarity over 202 aa) to the hypothetical ORF1 of the *Tennessee* element from *P. primaurelia* [34]. One of the *Sardine* copies (copy S6) is interrupted by the insertion, within the putative tyrosine recombinase gene, of a different but similar element, named *Thon* (French for "tuna"), which also contains a DD35E transposase, a tyrosine recombinase, possibly the two hypothetical ORFs, and palindromic TIRs of ~700 bp (Figure S5).

IES copy number and similarity to transposon sequences

For a handful of IESs, it has been shown experimentally that they are single copy elements [5]. In order to see whether this is generally the case, we looked for all IESs present in more than 1 fully identical copy (100% sequence identity). We found 44,210 IESs to be unique copy (98.4%). We examined all IESs present in 2 or more identical copies and found 39 cases of duplicate IESs as a result of errors in assembly of the MAC reference genome that had led to small, partially redundant scaffolds (4% of the MAC assembly is contained in scaffolds <45 Kb and some of these are partially redundant with the chromosome-size scaffolds [1]). The rest of the 319 IESs found in 2 copies were inserted in homologous genomic sites and appeared to be the result of recent segmental duplication or gene conversion. The 23 cases of IESs found in 3 to 6 copies correspond to expansion or recombination of repeated sequences such as tetratricopeptide repeat (TPR) domains or WD40 repeats.

We performed an all by all sequence comparison of the IESs and of their flanking sequences to see whether we could identify homologous IESs inserted at non-homologous sites in the genome. As shown in Table 2, we were able to identify 8 clusters of 2 to 6 IESs that share significant homology (BLASTN E-value $<10^{-10}$) over at least 85% of their length, inserted in non-homologous sites (cf. Text S2 for the alignments). Moreover, we found significant nucleotide identity (E-value 9×10^{-57} for the best match; nucleotide identity between 68 and 78% for the HSPs) between the IESs of cluster 5 and one of the Tc1/mariner-like transposons identified using the phage library (*Thon*, Figure S5). This is a strong indication that these IESs are derived from recently mobile elements.

However, the IES sequences of this cluster correspond to a single palindromic TIR. This might reflect assembly problems given use of a single insert size for the paired-end sequencing,

Table 2. Homologous IESs at non-homologous sites in the genome.

CLUSTER	IES SCAFFOLD	IES POSITION	SIZE (bp)	LOCATION	NUCLEOTIDE MATCH
2	scaffold51_25	381101	209	GSPATP00009750001	
	scaffold51_25	389332	213	intergenic	
3	scaffold51_117	944	608	intergenic	
	scaffold51_160	10020	577	intergenic	
	scaffold51_44	7711	555	intergenic	
5	scaffold51_109	40673	571	intergenic	TIR *Thon*
	scaffold51_128	266698	689	GSPATP00032295001	TIR *Thon*
	scaffold51_131	262422	630	intergenic	TIR *Thon*
	scaffold51_18	127217	770	GSPATP00007326001	TIR *Thon*
	scaffold51_34	280841	512	intergenic	TIR *Thon*
	scaffold51_58	302214	640	intergenic	TIR *Thon*
9	scaffold51_19	475992	666	intergenic	
	scaffold51_96	236752	665	intergenic	
12	scaffold51_124	248174	568	intergenic	
	scaffold51_27	275392	476	GSPATP00010339001	
13	scaffold51_155	211807	458	intergenic	
	scaffold51_20	46790	505	intergenic	
	scaffold51_27	294496	472	GSPATP00010351001	
14	scaffold51_184	21279	1024	GSPATP00038454001	
	scaffold51_21	430950	1038	GSPATP00008497001	
	scaffold51_58	200038	1010	GSPATP00018841001	
15	scaffold51_28	278632	262	GSPATP00010625001	
	scaffold51_4	361312	242	GSPATP00001801001	

A BLASTN internal comparison of all IESs, carried out with an E-value cutoff of 1e-10, was filtered for HSP coverage of at least 85% of the longest IES and for the absence of significant homology between 500 nt of MAC flanking sequence. The IESs were than transitively clustered and aligned using MUSCLE (Text S2). Some clusters were eliminated because of low complexity of the IES sequences. BLASTN homology searches at NCBI and against known Paramecium transposons ([34] and the present manuscript) were carried out using each IES in the clusters as query. *Thon* is a Tc1/mariner-like transposon. BLASTX similarity searches against the non-redundant protein database at NCBI did not yield any significant hits at an E-value cutoff of 0.001. The location of the IES, if in a coding sequence, is provided as a ParameciumDB accession number.

doi:10.1371/journal.pgen.1002984.t002

either because these IESs contain sequences repeated elsewhere in the genome or because the *Thon* TIRs are large (~700 bp) and palindromic so that the assembly might have jumped from one TIR to the other deleting the rest of *Thon*. We therefore used a long-range PCR strategy capable of amplifying large DNA fragments containing each of the IESs to verify their size and attempt to obtain sequences (detailed in Text S3). Amplification products of the expected sizes were obtained for all of the IESs from cluster 5, making it unlikely that these IESs correspond to a complete *Thon* element that had failed to be assembled from the paired-end sequencing reads. Three IESs were chosen for sequencing, and the sequences of the corresponding PCR products confirmed the IESs, indicating that they had been correctly assembled. Identification of 6 IESs (at non-homologous genomic sites) that share sequence identity with a *P. tetraurelia* Tc1/mariner solo TIR argues that at least a fraction of IESs do originate from Tc1/mariner-like elements.

We therefore adopted a complementary strategy, using the PFAM-A library of curated protein domains to search for domain signatures in the genome-wide set of IESs. Matches at a BLASTX E-value cutoff of 1 were inspected visually to filter out matches with PFAM-A protein domains from *Paramecium* and matches owing to compositional bias (high A+T content). This left 6 IESs, ranging in size from 2416 to 4154 bp, with a DDE_3 (PFAM accession number 13358) DDE superfamily endonuclease domain characteristic of IS*630*/Tc1 transposons. The peptides encoded by the IESs were subjected to an HMM search of the PFAM-A hmm profiles (http://pfam.sanger.ac.uk/search) for confirmation of the conserved residues and to validate the statistical significance of the match (E-values of 0.02 to 2.1×10^{-15} for the 6 peptides). The IESs were aligned with MUSCLE and a neighbor-joining tree grouped 4 of them together with good bootstrap values (not shown). The 4 IESs were used to search for sequence similarity with the genome-wide set of IESs and this allowed identification of 28 IESs ranging in size from 1251 to 4154 bp (Table S4). The IESs were aligned to provide the consensus sequence for 2 distinct Tc1/mariner-like 3.6 kb transposons from the same new family, baptized *Anchois* (Anchovy). Manually adjusted alignments used to reconstruct the *AnchoisA* and *AnchoisB* elements, consensus sequences and annotation are provided in Text S1.

Alignment of the DDE domains of the reconstituted *Anchois* transposons with the DDE domains from bacterial IS*630* elements, invertebrate Tc1 transposons and all known ciliate Tc1/mariner elements indicates that the *Anchois* transposase belongs to the IS*630*/Tc1 subfamily (Figure 2A). Unlike *Thon* and *Sardine* but like the *P. primaurelia Tennessee* element, *Anchois* TIRs are short and lack internal palindromes, moreover *Anchois* does not contain a putative tyrosine recombinase. *Anchois* has 2 hypothetical

ORFs in addition to the DDE transposase (Figure 2B; Text S1). The ORF2 of *Anchois* displays homology to ORF2 of *Sardine* (36.2% identity and 56.2% similarity over 210 aa) and to ORF1 of *Tennessee*. Interestingly, for 6 of the 28 IESs that initially identified the copies of *Anchois*, the *Anchois* TIRs do not correspond to the extremities of the IES, raising the possibility of *Anchois* insertions within pre-existing IESs. The discovery of the *Anchois* elements and the fact that several IESs appear to be full-length copies, provides a strong, direct link between IESs and transposons.

A remarkable IES size distribution

The size distribution of the genome-wide set of IESs is shown in Figure 3A, for the 93% of the IESs that are shorter than 150 bp. The most remarkable feature is a periodicity of ~10 bp, which corresponds to the helical repeat of double-stranded DNA. The first peak of the size distribution has maximal amplitude at 26–28 bp and includes 35% of all identified IESs. The abrupt cutoff at 26 bp represents the minimum IES size. A second peak appears to be forbidden and contains only a few IESs. The following peaks are centered at approximately 45–46, 55–56, 65–66 bp etc. and the distance between these peaks is best fit by a 10.2 bp sine wave (not shown). At the far end of the spectrum, 95 of the IESs are between 2 and 5 Kb in size. Similar periodic size distributions are found for IESs inserted in coding sequences and for IESs inserted in non-coding sequences (Figure S6). This indicates that the constraint on the distance between IES ends is

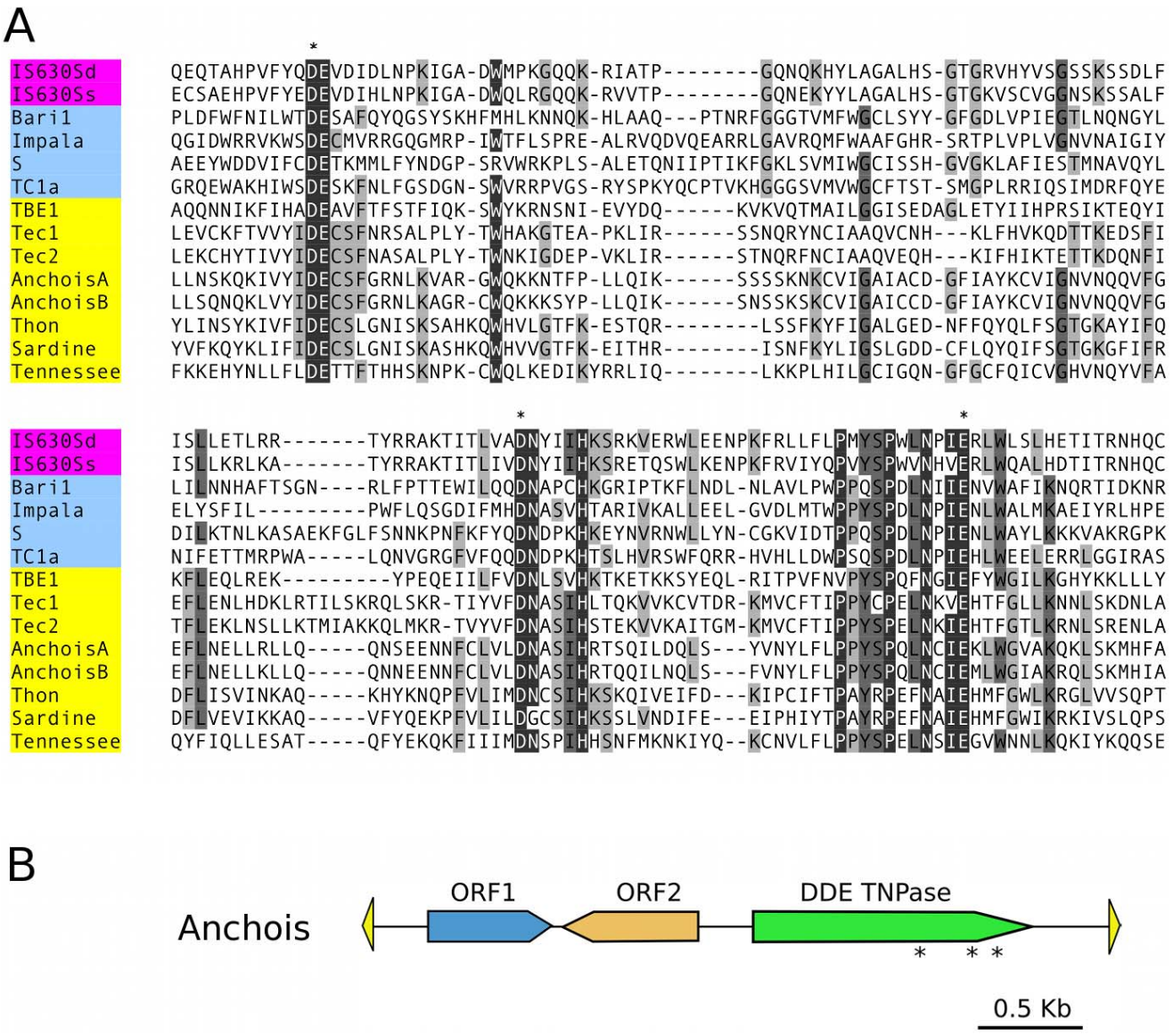


Figure 2. *Anchois* Tc1/mariner family transposon. A) Alignment of the DDE domains of bacterial IS*630* elements (IS*630*Sd, *Salmonella dublin*, GenBank Accession No. A43586; IS*630*Ss, *Shigella sonnei*, X05955), invertebrate and fungal Tc1 transposons (Baril, *D melanogaster*, Q24258; Impala, *Fusarium oxysporum*, AF282722; S, *D* melanogaster, U33463; Tc1, *C elegans*, X01005) and ciliate Tc1/mariner transposons (TBE1, *Oxytricha fallax*, L23169; Tec1 and Tec2, *Euplotes crassus*, L03359 and L03360; Anchois, Thon and Sardine, *Paramecium tetraurelia*, this article; Tennessee, *Paramecium primaurelia*, [34]). Asterisks mark the conserved catalytic DDE residues. B) Schematic diagram of the 3.6 Kb *Anchois* consensus, showing the position and orientation of the 3 ORFs. The yellow triangles represent the ~22 nt TIRs. Asterisks mark the position of residues of the catalytic DDE triad for the ORF encoding the DDE transposase.
doi:10.1371/journal.pgen.1002984.g002

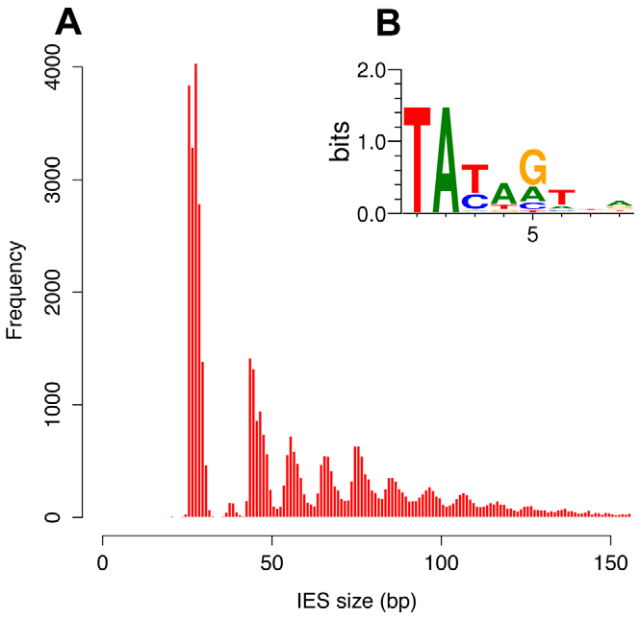


Figure 3. IES sequence properties. A) Histogram of the sizes of the genome-wide set of IESs that are shorter than 150 bp. B) Sequence logo showing information content at each position, corrected for a G+C content of 28%, for the ends of the genome-wide set of IESs.
doi:10.1371/journal.pgen.1002984.g003

an intrinsic property of the IESs and is not related to the locus in which they are inserted in the genome. Whatever their size, the IESs adhere to the weak, Tc1/mariner-like end consensus first reported for 20 IESs located in surface antigen genes [18], as illustrated in Figure 3B for the whole set. Differently sized subsets of the IESs all display essentially the same end consensus (data not shown).

We further examined constraints on IES size and sequence by evaluating IES conservation with respect to the 3 WGDs in the *Paramecium* lineage. We used the large number of paralogs (hereafter termed "ohnologs") of different ages (Table 3) that could be identified for each of the WGD events [1] to ask whether IESs are present, at the same position relative to the gene coding sequences, in ohnologs of the different WGD events. This analysis makes the assumption that IES insertions are rare events so that if IESs are present at the same position in ohnologous genes, then they must have been acquired before the WGD and can be considered to be "ohnologous" IESs. As shown in Table 3, we found 84.5%, 23.2% and 5.9% conservation of IESs with respect to the recent, intermediate and old WGDs respectively. For comparison, more than 99% intron conservation was found for 1,112 pairs of genes related by the recent WGD [35]. This indicates that the dynamics of IES insertion or loss over evolutionary time is relatively fast compared to that of introns. The only phylogenetic study of IESs, carried out for two loci in a few different stichotrich (formerly called hypotrich) ciliates, which are very distantly related to *Paramecium*, also concluded that the intragenic IESs in those species evolve very rapidly [36]. We found that the ohnologous IESs related by the recent WGD are highly divergent in sequence. In more than 90% of cases, the sequence identity was too low for detection by BLASTN (E-value threshold of 10^{-5}). This high level of sequence divergence is consistent with the pattern expected for neutrally-evolving non-coding regions, since the average synonymous substitution rate measured between ohnologous genes derived from the recent WGD is about 1 substitution per site [1]. However, if we compare the lengths of IESs that are conserved with respect to the recent WGD (Figure 4A), for ~55% of the pairs, both IESs are found in the same peak of the IES size distribution. The honeycomb appearance of the plot (Figure 4A), with off diagonal cells that result from ohnologous IESs in different peaks of the distribution, underscores the strong evolutionary constraint that is exerted on IES size.

Table 3. IES conservation in ohnologs produced by the different WGDs.

WGD event	Genes with ohnolog	IESs	Conserved IESs	% conserved
Recent	24052	20623	17430	84.5
Intermediate	12590	11561	2675	23.2
Old	3381	3646	215	5.9

The identification of ohnologs and the reconstitution of the pre-duplication genomes is described in [1]. For the most recent WGD, which preceded the appearance of the *P. aurelia* complex of 15 sibling species [95], 51% of the pre-duplication genes are still present in 2 copies. For the intermediate duplication, 24% of pre-duplication genes are still present in 2 or more copies. For the ancient duplication, which preceded the divergence of *Paramecium* and *Tetrahymena*, 8% of pre-duplication genes are still present in 2 or more copies. The significance of the column headers is as follows. Genes with ohnolog: the number of present day genes with at least one ohnolog from the indicated WGD event. IESs: the number of IESs found in the genes with at least one ohnolog from the indicated WGD event. Conserved IESs: number of IESs found at the same position in at least one other ohnolog, as determined by sequence alignment. The identification of ohnologs is described in [1] and the data are available through ParameciumDB [90]. Note that this analysis only concerns IESs that are within paralogous genes and not IESs found in intergenic regions.
doi:10.1371/journal.pgen.1002984.t003

Dynamics of IES gain and loss

In order to investigate the rate of IES insertions and losses during the evolution of the *Paramecium* lineage, we examined gene families, which we call "quartets", for which all 4 ohnologs issued from duplication of an ancestral gene at the intermediate and then the recent WGD are still found in the present day genome. Of the 1350 such quartets identified in the MAC genome [1], 878 contain at least one IES in at least one of the 4 duplicated genes. We evaluated the conservation of IESs at the same position with respect to the coding sequence for all members of each quartet (Figure S7), and identified 2126 IES groups, each group containing an IES conserved either in all 4 genes ($N_{1111} = 190$), in 3 genes ($N_{1110} = 64$), in 2 genes on the same intermediate WGD branch ($N_{1100} = 1304$), in 2 genes on different branches ($N_{1010} = 10$) or in only one of the 4 genes ($N_{1000} = 558$).

Under the assumption that two IESs present at the same location in ohnologous genes derive from a single ancestral IES (i.e. the probability of two insertion events occurring at the same site after a WGD is considered negligible), and that the rate of IES losses has remained constant, it is possible to estimate the rate of IES gain during the evolution of the *Paramecium* lineage (the model is developed in Text S4). The quartet analysis is fully consistent with a model whereby IES acquisition has been ongoing since before the intermediate WGD (15% of the IESs predating this WGD), with a peak in the period between the intermediate and the recent WGD events: 69% of IESs were acquired during the interval between these two WGDs, vs. 16% during the period since the recent WGD, which corresponds to about the same evolutionary time. Genome-wide IES data for other *Paramecium* species will be necessary in order to test the assumption of a

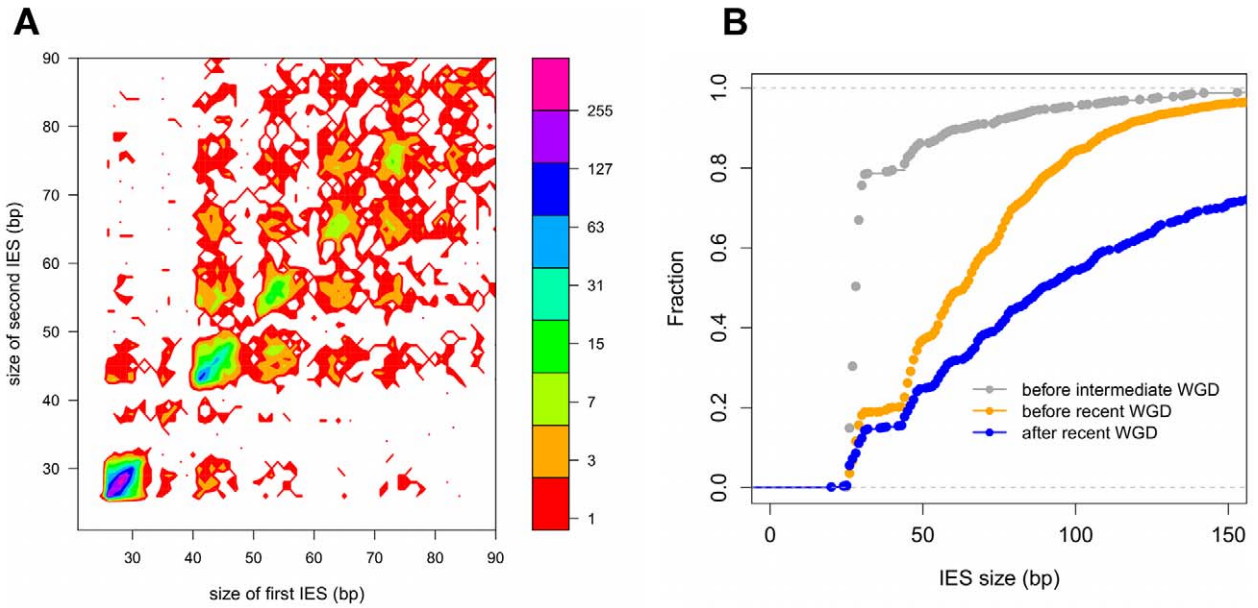


Figure 4. IES conservation in genes related by WGD. A) Filled contour plot of the correlation between the size of IES pairs that have been conserved with respect to the recent WGD. The x axis gives the size in bp of the first IES, the y axis gives the size in bp of the second IES found in the ohnologous gene and the color of each point indicates the number of times that combination of x,y values was found in the data set. The color legend is shown to the right of the figure, the numbers represent counts of the x,y value pairs; the rainbow colors are distributed according to a log2 scale. B) Size distribution of IESs conserved in "quartets" i.e. genes that are still present in 4 copies in the genome after duplication at both the intermediate and the recent WGD events. In order to compare size distributions for different classes of IES, they are represented as experimental cumulative distribution functions. The ripples in each curve correspond to the peaks of a histogram representation as in Figure 3A. The curves are for IESs that must have originated from an ancestral IES acquired before the intermediate WGD (grey, N_{1111} IESs), IESs that must have originated from an ancestral IES acquired before the recent WGD (orange, N_{1100} IESs) and the IESs that might have been acquired since the recent WGD (blue, N_{1000} IESs).
doi:10.1371/journal.pgen.1002984.g004

constant rate of IES losses. However, even if we relax this assumption (i.e. rates of IES losses are allowed to vary over time), the model still strongly rejects the hypothesis that all IESs were acquired before the intermediate WGD (cf. Text S4). Thus, with the presently available data and biologically reasonable assumptions, we conclude that IESs have been acquired in all 3 of the time periods delimited by the intermediate and recent WGD events.

We compared the cumulative size distributions of the N_{1111}, N_{1100} and N_{1000} IESs (Figure 4B). The N_{1111} IESs, which must have been acquired before the intermediate WGD, are much shorter than the IESs of the two other samples, with almost 80% of the IESs in the first peak, compared to 20% for N_{1100} IESs, which may mainly result from IES acquisition after the intermediate but before the recent WGD, and only 16% for N_{1000} IESs, at least some of which may have been acquired since the recent WGD. In addition, the curves are significantly shifted with respect to each other, in particular, 30% of the N_{1000} IESs are larger than 150 nt, compared to scarcely any IESs larger than 150 nt for the two other samples. This analysis shows that the older an IES, the shorter it is likely to be, consistent with a decay process involving progressive shortening of IESs by accumulation of small deletions, in addition to the accumulation of point mutations.

Quartet analysis is restricted to IESs in genes that have been retained in 4 copies (fewer than 10% of all IESs). Similar distributions of IES size are found if we consider all ohnologous IESs (45% of all IESs, cf. Table 3). IESs conserved with respect to the intermediate WGD (76% of IESs in first peak) are significantly shorter than IESs conserved only with respect to the recent WGD (30% of IESs in the first peak) (data not shown). The size distribution of IESs conserved with respect to the old WGD is poorly determined because of the small number of conserved IESs (Table 3), which are moreover often in genes that have undergone recent gene conversion judging from the nucleotide divergence of the ohnologs (data not shown). It is therefore uncertain that IESs were present in the genome before the old WGD, consistent with the absence of TA-bounded IESs in *Tetrahymena*, which diverged from *Paramecium* after the old WGD event [1].

Since we found essentially no IESs shorter than 26 bp, it seems likely that some mechanism(s) other than decay of the sequence through internal mutations and deletions is responsible for the complete loss of an IES. In order to explore this question, we examined case by case, using both nucleotide and conceptual protein alignments, all of the N_{1110} quartet IES groups (n = 64), which are most parsimoniously explained by insertion of an IES before the intermediate WGD followed by loss of an IES after the recent WGD. We examined the raw read alignments and PGM and phage contigs in order to be sure that there was sufficient read coverage and no evidence suggesting presence of an IES at any site of putative IES loss. We found 4 different explanations for the quartet triplets: precise loss of the fourth IES (n = 17), gain of the third IES by gene conversion between intermediate WGD ohnologs (n = 1), recruitment of the fourth IES into the exon sequence (n = 6), and deletion of the region that encompasses the fourth IES (n = 23), often testifying to the formation of a pseudogene. In addition, we found 5 errors in IES detection (the fourth IES probably exists as it can be found in the phage contigs or is predicted by the MIRAA pipeline). In the remaining cases (n = 12), annotation or alignment problems made it difficult to conclude. The observation of 17 cases of precise loss of an IES from the germline DNA raises the possibility that there is a mechanism for conversion of a MIC locus to the IES-free form

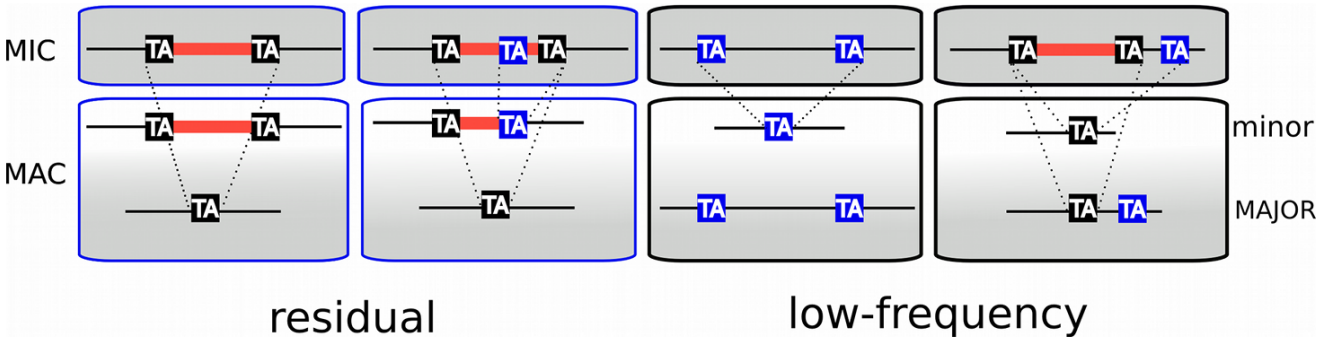


Figure 5. TA-indels are produced by IES excision errors. Schematic representation of the "residual" and "low frequency" TA-indels that were identified by comparing the MAC draft genome assembly (MAJOR form) with the 13× Sanger sequencing reads used to build the assembly [29]. The TA-indels were identified by one or more reads that differed from the assembly (minor form). The residual TA-indels were assumed to be the result of occasional failure to excise an IES and the low-frequency TA-indels to result from excision of MAC-destined sequences. Comparison of the genome-wide set of IESs with the TA-indels revealed that many TA-indels result from the use of alternative IES boundaries situated inside the corresponding IES in the case of residual TA-indels and outside the IES in the case of low-frequency TA-indels. In the schema, TA dinucleotides in black boxes are *bona fide* IES boundaries while TA dinucleotides in blue boxes are alternative IES boundaries.
doi:10.1371/journal.pgen.1002984.g005

using a MAC genome template. However, we cannot rule out the possibility that IESs can be precisely excised from the MIC DNA, and therefore lost, by the same Pgm-dependent mechanism as that involved in MAC genome assembly.

TA-indels reveal IES excision errors

The analysis of sequence variability in the polyploid (800n) MAC genome, carried out by comparing the MAC assembly representing a "consensus" sequence with the 13× Sanger sequencing reads used to build the assembly, revealed nearly 2000 "TA-indels" that were presumed to be produced by the IES excision machinery and to reflect excision errors [29]. As shown schematically in Figure 5, "residual" TA-indels (n = 739), that were suggested to represent occasional retention of IESs on some macronuclear copies, were absent from the assembly ("major" form in Figure 5), but present in at least one sequence read ("minor" form). For 689 of the residual TA-indels (93%), we found an IES at the corresponding site in the genome. Interestingly, in 134 cases (19.4%), the TA-indel was shorter than the IES and case by case inspection indicated that most of these TA-indels may be products of IES excision that used an alternative IES boundary located within the IES (Figure 5). In this case, the TA-indel would only correspond to part of a larger IES. A few cases of use of an alternative IES boundary that may confer a mutant phenotype have been reported [7,37].

"Low frequency" TA-indels (n = 1090), previously suggested to represent excision of MAC-destined sequences [29], were present in the assembly (major form, Figure 5), but absent from at least one sequence read (minor form). We could not look for the "low-frequency" TA-indels directly among the genome-wide set of IESs, since they are part of the MAC genome assembly. However, we examined the ends of the low-frequency TA-indels and found 249 cases (23%) where the TA dinucleotide at one of the ends corresponds to the insertion site of an IES in the genome-wide set (Figure 5), indicating that the TA-indel was generated by use of an alternative IES boundary located outside of the IES. The whole of the analysis supports the previous conclusion [29] that TA-indels are products of the IES excision machinery. The high incidence of alternative boundaries in both classes of TA-indels, revealed by comparing them with the genome-wide set of IESs, strengthens the previous conclusion [29] that TA-indels reflect IES excision errors (see below). Thus TA-indels cannot be considered to be IESs in the absence of further experimental support.

Evidence for selective pressure against IES insertion

IESs are tolerated in coding sequences and evolve under a strong constraint on their size and end-consensus, properties that are presumably important for their precise and efficient excision. However, the excision machinery can commit errors, as revealed by TA-indels (cf. above) and by the use of alternative IES

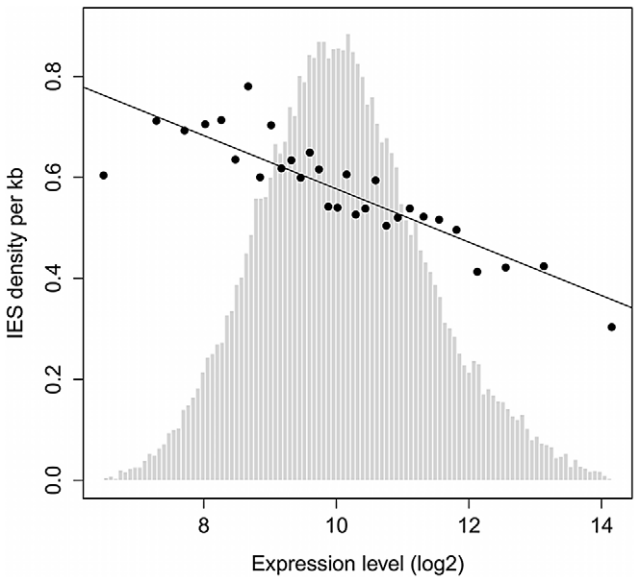


Figure 6. IES density is inversely proportional to gene expression level. Genes were binned according to their median expression level across 58 microarrays representing different cellular and growth conditions as described in [38,39]. The expression levels were divided into 30 bins as in [38]. The black points show the average IES density (per Kb) of genes in each bin. Linear regression was used to fit the points. Light gray bars show the distribution of genes according to their expression level (before binning).
doi:10.1371/journal.pgen.1002984.g006

Table 4. Deficit of 3n IESs in coding sequences.

IES Category	Number	3n	non-3n	χ^2	*P*-value
Non-coding	10304	3481 (33.78%)	6823 (66.22%)	-	-
Coding stopwith	11205	3700 (33.02%)	7505 (66.98%)	2.91	0.08
Coding stopless	23339	7095 (30.40%)	16244 (69.60%)	119.42	8.47×10^{-28}
Q1 stopless	6044	1892 (31.30%)	4152 (68.70%)	16.61	4.59×10^{-5}
Q4 stopless	5712	1615 (28.27%)	4097 (71.73%)	77.5	1.32×10^{-18}

For the calculation of χ^2, the observed numbers of IESs of length 3n and non-3n inserted in coding sequences are compared to the distribution found for IESs inserted in non-coding sequences under the null hypothesis that IES length is not under constraints related to translation. The null hypothesis is rejected only for those IESs inserted in coding sequences that do not contain a stop codon in frame with the upstream ORF (Sample "Coding stopless"). Microarray experiments [38] were used to group the IESs according to the expression level of the genes in which they are inserted. "Q1" designates IESs in exons of the 25% least expressed genes and "Q4" designates IESs in the exons of the 25% most expressed genes, those subject to the strongest selective pressure. The bias against 3n IESs is stronger in the Q4 sample than in the Q1 sample.
A more detailed analysis of the modulo 3 length distribution for IESs in coding and non-coding sequences, for each peak of the 10 bp periodic size distribution, is provided in Table S5.
doi:10.1371/journal.pgen.1002984.t004

boundaries [7,37]. We therefore looked for evidence that the rate of excision errors is high enough to represent a fitness burden for the organism. First, only 47% of genes contain at least one IES, and the IESs are less represented in strongly expressed genes. Figure 6 shows the density of IESs in genes as a function of gene expression level determined by microarray experiments [38,39]. The density varies from about 0.7 IESs per Kb (i.e. an IES on average every 1.4 Kb) in genes with low expression to less than 0.3 IESs per Kb (i.e. an IES on average every 3.3 Kb) for the genes with the highest expression. The inverse correlation observed across all levels of expression indicates that IESs are less-well tolerated the more a gene is expressed.

Second, IESs inserted in protein-coding exons display a characteristic bias in their size. There is a statistically significant deficit in IESs whose length is a multiple of 3, compared to IESs found in non-coding regions. Furthermore, this bias is only found for 3n IESs that do not contain a stop codon in phase with the ORF of the upstream coding sequence (Table 4; cf. Table S5 for a more detailed analysis). A similar 3n bias was reported for introns in eukaryotic genomes, and experiments in *Paramecium* showed that the Nonsense Mediated Decay (NMD) pathway destroys mRNAs containing unspliced introns, provided the intron retention leads to a premature stop codon [35]. Retention in mRNA of a 3n stopless intron would not be detected by NMD and therefore could lead to translation of potentially harmful proteins, explaining the deficit in 3n stopless introns. The fact that IESs display a similar deficit suggests that the rate of IES retention is high enough to represent a fitness cost, so that IESs in exons are under selective pressure to be detected by NMD in case they are retained in the MAC genome. We were able to test this hypothesis by looking at the size bias for IESs located in the exons of the 25% of *Paramecium* genes that are the most highly expressed hence subject to the strongest selective pressure. As shown by the last 2 lines of Table 4 (samples Q1 and Q4), the deficit in 3n IESs is the greatest for the IESs found in the most highly expressed genes (28.3%), where IES retention would be the most deleterious.

Discussion

An IES reference set for *P. tetraurelia*

Previous studies of *Paramecium* IESs all relied on a small reference set of about 50 IESs. For the first time in any ciliate genome, in so far as we are aware, we have carried out an exhaustive identification of IESs. Since it is not yet possible to isolate *Paramecium* MICs in the quantity and of the purity required for genomic sequencing, we relied on nuclear DNA isolated from cells depleted in Pgm, the domesticated transposase required for introduction of the DSBs that initiate IES excision [21]. We fortunately were able to use the only genomic library ever made from purified MICs [28] – but heavily contaminated by bacterial DNA – to obtain genome-scale evidence that Pgm is required for excision of all *Paramecium* IESs and to estimate that our IES reference set includes ~98.5% of all IESs.

Although this IES reference set will prove useful for a variety of studies, it is important to keep two things in mind. First, the IES definition used here is necessarily a genomic definition involving comparison of MIC and MAC sequences. Our procedure does not allow identification of nested IESs (unless the external IES is retained in the MAC), or of any IES located in part of the MIC genome that is not collinear with MAC chromosomes. The complexity of the assembled PGM DNA is almost 100 Mb, although we could not properly assemble repeated sequences. We thus estimate that at least 25% of the germline is not collinear with the MAC chromosomes, and might contain unique copy IESs or transposons, the excision of which could only have been detected if the flanking region were retained in the MAC.

Second, this reference set does not provide information about the variability in IES excision patterns that might exist between different, though genetically identical, cell populations. Many IESs are under maternal, epigenetic control [40,31,41]. The genome scanning model [42] posits that every time *Paramecium* undergoes meiosis, the scnRNA pathway compares the maternal MIC, in the form of 25 nt scnRNAs [43], with the maternal MAC, in the form of long non-coding transcripts [44]. The scnRNAs that cannot be subtracted by base pairing with the long maternal transcripts are licensed for transport into the new developing MAC [45] where they target homologous sequences for elimination, probably via deposition of epigenetic marks on the chromatin (cf [3,4] for recent reviews of genome scanning in *Paramecium* and *Tetrahymena*). The scnRNA pathway in theory provides a powerful defense mechanism against transposons that invade the germline and can explain the molecular basis of alternative MAC rearrangement patterns that are maintained across sexual generations [31,40,41,46,47]. Hence the following caveat: any genome-wide set of IESs is identified with respect to a particular MAC reference genome sequence. There can be no "universal" IES reference set for the species. Since IESs can be a source of genetic variation as discussed in [48], the IES catalogue we have established will make it possible to study this variation, for example by surveying IES retention in the MACs of geographic isolates and in stocks that have been experimentally subjected to different types of stress.

Constrained IES size distribution and the IES excision complex

The remarkable sinusoidal distribution of IES sizes retained by evolution reflects strong constraint on the distance between IES ends. We assume that the selection is exerted through the excision mechanism, since the retention of an IES in the MAC can impair gene function. An IES that cannot be efficiently excised is expected to be counter-selected. We propose an interpretation of the IES size distribution based on its similarity with data generated

by "helical-twist" experiments, which have provided evidence of DNA looping between distant protein-binding sites in various, mainly prokaryotic, DNA transaction systems (transposition, gene control, replication initiation, site-specific recombination, etc. reviewed in [49]). In these experiments, the distance between transposon ends [50,51], repressor binding sites [52–55] or site-specific recombination sites [56] is varied, on plasmids or on the bacterial chromosome, and the activity of the system is measured *in vivo*. The observed periodicity in the length-dependence of the activity corresponds to the helical repeat of the DNA, since the same face of the double helix must interact with the protein at each end, and given the prohibitive energetic cost of twisting the double helix to fit the binding site to the protein. This is especially true for DNA fragments whose size is close to the persistence length of double stranded DNA (~150 bp) or shorter. The persistence length, a physical measure of the bending stiffness of a polymer in solution, is the length above which there is no longer a correlation between the orientation of the ends of the molecule. For DNA longer than its persistence length, it becomes possible for the 2 ends to encounter each other to form a loop, without any external intervention.

Almost all (93%) of the IESs in the genome are shorter than the persistence length of DNA. The size distribution, which appears as a series of regularly spaced peaks, can be decomposed into three parts. The largest peak is centered on 28 bp but displays an abrupt minimum size cutoff at 26 bp. A second peak seems to be of forbidden size. Finally, there follow a series of peaks that are best fit by a sine wave with a ~10.2 bp periodicity. In the helical-twist experiments, the amplitude of the measured biological activity peaks tends to decrease with decreasing distance between interacting sites. However, for the IES size distribution, the decay of IESs over time imposes the opposite tendency: the peak heights increase as IES size decreases.

Our working model for assembly of an active IES excision complex is shown in Figure 7. We propose that, starting at the third peak (44–46 bp), the IESs assemble into the excision complex by forming a double-stranded DNA loop compatible with presentation of the same face of the double helix to the Pgm endonuclease at both IES ends. The near absence of the second peak, the minimum IES size of 26 nt and the 13 bp size of each *piggyBac* TIR [25] lead us to suggest that the IESs in the first peak are able to assemble an active excision complex without formation of a DNA loop. The IESs in the nearly absent second peak would not be efficiently excised, as they would be too short to form a DNA loop and too long to form an active excision complex without a DNA loop.

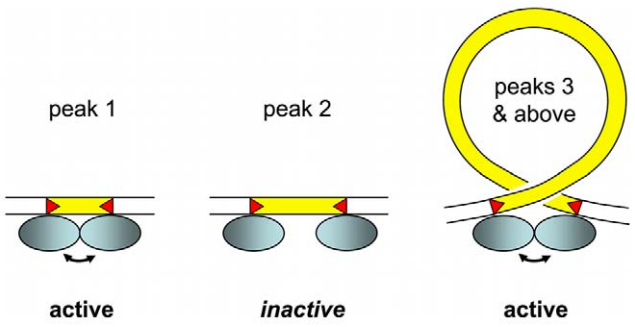


Figure 7. IES size constraint and the assembly of an active excision complex. Our working model is based on the assumption that oligomerization of the IES excisase (most likely the domesticated transposase PiggyMac) on DNA activates catalytic cleavage at IES ends (IESs are drawn in yellow and red triangles highlight the orientation of their ends). In the absence of any information on the stoichiometry of the complex, the excisase is represented by a shaded blue ellipse. For very short IESs from peak 1 (26–30 bp in length), the required contact between protein subunits may be established directly (double-headed arrow) and the complex is active. For IESs longer than 44 bp (peak 3 and above), we propose that looping of the intervening DNA double helix brings IES ends into close proximity and activates DNA cleavage. We have arbitrarily drawn the complex as an antiparallel arrangement of IES ends within a negatively supercoiled loop, but other conformations are possible. IESs from the "forbidden" peak 2 would be too long to allow direct contacts between protein subunits to be established, and too short to form an excision loop.
doi:10.1371/journal.pgen.1002984.g007

Molecular analysis of the IES excision mechanism supports the involvement of such a transpososome-type excision complex. First, the domesticated Pgm transposase, which has retained the catalytic site of *piggyBac* transposases [21], is very likely to be the endonuclease responsible for the cleavage reaction, involving the introduction of DSBs at each end of the IES [24]. Second, for IESs larger than 200 bp, covalently closed circular molecules containing the excised IES have been detected as transient intermediates during MAC development [57]. Third, if one end of an IES bears a mendelian mutation in the TA dinucleotide, no DSB occurs at either end of the IES. This indicates that the two IES ends must interact, directly or indirectly, before cleavage can occur [58].

It is worth noting that "canonical" TIRs of cut-and-paste transposons are often bipartite. They are composed of an internal sequence motif recognized and bound by the transposase, and of a few nucleotides at the termini that constitute the DNA cleavage site [59]. The obligatory conservation of a TA dinucleotide at IES ends is indicative of a requirement for DNA cleavage but is not sufficient for specific recognition, even if we take into account the weak consensus over the 6 internal nucleotides. The lack of a sufficiently long conserved motif in IESs makes it unlikely that Pgm recognizes IESs by binding to a specific sequence. For IESs under maternal control [31], it is currently thought that Pgm is recruited to its substrate via epigenetic marks deposited on the chromatin by the scnRNA pathway [3,21,42].

The picture of an IES excision complex that emerges from these considerations, which must of course be tested biochemically, requires very short pieces of DNA to form loops (Figure 7). Proteins that bend DNA, such as HMG proteins [60], could be involved. What is quite remarkable here, beyond the fact that evolution has performed such a nice "helical-twist" experiment, is that the DNA loops might be as short as ~45 bp, shorter than almost any reported case of DNA looping. The minimal *in vivo* value reported for cut-and-paste bacterial transposons is 64–70 bp [50,51] and this is also the minimum size reported for HMG assisted DNA loop formation *in vitro* [60]. The only indication of shorter loops comes from detection of a minor peak of activity *in vivo* and *in vitro* for ~50 bp DNA loops in the *E. coli* Hin invertasome, provided that invertasome assembly occurs in the presence of HU, a bacterial nucleoid protein that bends DNA [56]. Given the unusually high A+T content of IESs (80%), local melting might favor the deformations in the double helix required to make the very small looped structures of the postulated IES excision complex.

Evidence that IESs are remnants of transposons

Ciliate MICs have long been recognized as safe havens for transposons, since removal of the transposons from the somatic DNA during development would decrease the burden on host fitness, as discussed in [19]. Our study provides the first global vision of IESs in any ciliate germline and provides strong support for the "transposon link" hypothesis that present day IESs are remnants of transposons [18,19].

Although we do not yet have a complete picture of the transposon landscape of the *P. tetraurelia* germline genome, we have identified 3 families of Tc1/mariner elements, with 2 quite different structures. The *Thon* and *Sardine* transposons have long, palindromic TIRs, a tyrosine recombinase and a DDE transposase characteristic of the IS*630*/Tc1 subfamily, with a short spacer (32 aa) between the 2nd and 3rd catalytic residues. This clearly distinguishes these transposons from the *piggyBac* family characterized by a long spacer and a DDD catalytic triad. The IESs related to these elements that we were able to identify appear as solo TIRs. Given the presence of repeated, palindromic subsequences in each TIR, we can speculate that the solo TIRs result from recombination between short direct repeats present within the complex TIRs, as proposed to explain the incidence of solo LTRs derived from LTR retrotransposons in the genomes of some organisms [61,62]. The other transposon family we have identified, *Anchois*, is characterized by much shorter TIRs which do not contain internal palindromes, a similar DDE transposase and the absence of a tyrosine recombinase. This structure is similar to that of the *P. primaurelia Tennessee* transposon [34]. In the case of *Anchois*, we could find a number of IESs that appear to correspond to the entire transposon or large portions of it, including IESs with a recognizable but degenerate DDE transposase ORF.

It is possible that we have only scratched the tip of the iceberg since the germline genome is expected to contain other mobile elements. Indeed, we were able to identify 8 clusters of homologous IESs inserted at non-homologous genomic sites, suggesting recent mobility, and one of these clusters turned out to consist of IESs that are solo TIRs of the *Thon* element. The other clusters could be the remains of as yet unidentified elements. Both the *Thon* and the *Anchois* IES homologies were detected among the largest IESs in the genome-wide set (i.e. the 380 IESs >500 bp), and for none of them could we detect ohnologous IESs from the recent WGD, an indication that these IESs were recently acquired. Since over 90% of present day IESs have decayed to very short sizes (<150 bp) it is not surprising that internal transposon motifs can no longer be recognized. These very short IESs nonetheless display the short degenerate Tc1/mariner end consensus. The existence of this consensus at IES ends may testify to their evolutionary transposon origin. This end consensus would eventually have become a requirement for efficient cleavage by the IES excision machinery. We can imagine two instances of such convergent evolution: i) other families of mobile elements could be eliminated by the PiggyMac-dependent mechanism and ii) genomic sequences that adhere to the end consensus could be excised just like IESs. We conclude that at least a fraction of IESs are decayed Tc1/mariner transposons, and we consider highly probable that some IESs are derived from other mobile elements.

IESs are a burden for host fitness

Since IES excision is not 100% efficient, IES insertions are in general deleterious, consistent with the different kinds of selective pressure we have observed: (i) a constrained IES size distribution likely reflecting assembly of the excision complex; (ii) a bias against IESs that do not lead to premature stop codons in case of IES retention in the MAC; (iii) an inverse correlation between IES insertions and gene expression level. IESs can in addition be considered to constitute a mutational burden, in the same way as introns are considered to constitute a mutational burden in intron-rich eukaryotic genomes [63], since IESs are present in large number in *Paramecium*, and any mutation in a flanking TA dinucleotide abolishes IES excision. Nonetheless, the system can give rise to beneficial new functions, as attested by use of the IES excision machinery to provide a regulatory switch for mating type determination (D. Singh, personal communication).

Since IESs are in general deleterious and constitute a fitness burden for the organism, and since we have detected cases of probable clean IES loss from the germline DNA suggesting that a mechanism exists for precise IES excision in the MIC, we may ask why *Paramecium* has any IESs at all. This question can be easily answered if we consider that IESs arise from selfish genetic elements (SGEs, defined as elements – typically transposable elements or viruses – that can enhance their own transmission relative to the rest of the genome, with deleterious or neutral effects for the host [64]). The number of IESs reflects the balance between the number of IES insertions (e.g. invasion by SGEs that subsequently decayed to become unique-copy IESs) and the strength of selection against these insertions, which either prevents fixation of new insertions in the population or favors loss of already fixed insertions. This genetic conflict is mediated by an "arms race" between SGEs and the host as discussed by Werren [64].

Host defense mechanisms in ciliates

In all kingdoms of life, non-coding RNAs are used to defend host genomes against parasitic nucleic acids, as exemplified in eukaryotes by small RNA pathways involved in protection against viruses or in silencing transposons to ensure integrity of the germline genome [65–67]. In ciliates, nuclear dimorphism provides the potential for an additional layer of protection by physically separating the chromosomes that store the genetic information from the rearranged chromosomes that express the genetic information. Additional host defense machinery providing precise excision of transposons/IESs from somatic DNA, might have allowed the invasion of a fraction of the genome in which SGEs are not usually tolerated, namely the coding and regulatory sequences required for gene expression.

In the case of *Paramecium*, Pgm domestication has provided the mechanism for precise excision of TA-bounded insertions from the somatic DNA, allowing transposons/IESs to be cleanly excised from genes in the MAC. Since this would reduce the fitness burden caused by transposition, we presume that it allowed transposons to spread throughout the MIC genome. Recognition of the IESs is however ensured by the scnRNA pathway [3], itself an example of the more ancient mechanism of small RNA-based host immunity against foreign nucleic acids, and this epigenetic recognition may in part explain the less than 100% efficiency of IES excision.

In *Tetrahymena*, which has both a scnRNA pathway and domesticated *piggyBac*-like transposases [4,22], only excision of intergenic IESs has been studied for the moment and use of heterogeneous cleavage sites was found. This imprecise excision would not be compatible with insertion in genes since gene expression would be compromised. *Tetrahymena* has only about 6,000 IESs and indeed, they are not usually found within genes [17]. Why doesn't *Tetrahymena* have intragenic IESs? We can only speculate that a Tc1/mariner invasion after the divergence of *Paramecium* and *Tetrahymena* was instrumental in the evolution of a precise excision mechanism in *Paramecium*, necessary for spread of these elements throughout the genome. In support of this hypothesis, a recent genome-scale identification of hundreds of *Tetrahymena* IESs [17] revealed a new class of TTAA-bound IESs that are precisely excised. They were found to contribute 3′ exons to genes that are expressed from the zygotic genome during genome rearrangements. These elements might be derived from *piggyBac* transposons, which have TTAA target sites, and perhaps

testify to the ancient *piggyBac* invasion that led to domestication of the transposase.

A contrasting situation is found in some stichotrich ciliates. The stichotrich ciliates are very distantly related to the oligohymenophorean ciliates and are characterized by highly fragmented somatic genomes consisting of nanochromosomes that usually bear a single gene. Intragenic IESs are more abundant in the germline genomes of *Oxytricha* and related strichotrichs than in *Paramecium*, with an estimate of at least 150,000 IESs per haploid genome [68]. Both single-copy IESs and transposons are precisely excised and the precise IES excision is assured by guide RNAs transcribed from the maternal MAC [69], which are even capable of reordering the scrambled MAC-destined gene segments that occur frequently in *Oxytricha* and related stichotrichs [70]. There is also evidence that the endonuclease required for cleavage in *Oxytricha* is actually a transposase from germline TBE transposons [71]. However, there is currently no evidence for a scnRNA pathway specialized in the control of DNA elimination, although gene silencing by RNAi in *Oxytricha* testifies to the presence of small RNA machinery [69]. Thus the high precision and fidelity of the guide RNA mechanism for genome rearrangements in *Oxytricha* spp. seems to have tipped the balance even further in favor of intragenic IES insertions.

The case of *Euplotes*, a stichotrich ciliate distantly related to *Oxytricha* and probably lacking scrambled genes, merits special attention. Beautiful work carried out by the Jahn and Klobutcher labs in the 1990s showed (i) the existence of high copy number Tc1/mariner elements, Tec1 (2,000 copies per haploid genome) and Tec2 (5,000 copies), as well as lower copy number Tec3 elements (20–30 copies) [33,72,73]; (ii) at least a fraction of these Tec elements are precisely excised between TA dinucleotides [74]; (iii) an estimated 20,000 short TA-bounded IESs [33], bearing a Tc1/mariner end consensus just like the *Paramecium* IESs [18], are excised precisely between TA dinucleotides leaving a single TA at each excision site on the MAC destined chromosomes [33] and (iv) molecular characterization of excised circular forms of both Tec elements and short IESs revealed an unusual junction consisting of 2 TA dinucleotides separated by 10 bp of partially heteroduplex DNA, showing that both the Tec transposons and the short IESs are excised by the same mechanism [74,75]. The mechanism is moreover different from that of precise IES excision in *Paramecium* [24,57]. Neither the endonuclease responsable for IES cleavage nor the repair pathway has currently been identified in *Euplotes*. It will be fascinating to see whether the same actors, i.e. a domesticated *piggyBac* transposase and the NHEJ (non-homologous end-joining) pathway, are responsible for a mechanism that in its details is not the same as that found in *Paramecium*, or whether completely different cellular machinery has been recruited to carry out the same function i.e. the precise excision from somatic DNA of the Tc1/mariner family Tec transposons and of short TA-bounded IESs presumed to be their relics [19].

In conclusion, different ciliates have evolved different host defenses in response to germline SGE insertions. In all cases that have been examined at the molecular level, maternal non-coding RNAs are involved in programming genome rearrangements. In *Paramecium* and some other lineages, the co-evolution of host defense machinery and SGEs has provided mechanisms for precise somatic excision, uniquely allowing the colonization of coding sequences by Tc1/mariner and likely other transposable elements. This phenomenon is so far only paralleled by the spread of introns into eukaryotic coding sequences, also thought to result from domestication of precise excision machinery, derived in this case from mobile self-splicing ribozymes [76].

Materials and Methods

Purification of DNA enriched in un-rearranged sequences from isolated nuclei of cells depleted for PiggyMac

Cell growth and autogamy. *Paramecium tetraurelia* strain 51 was used for this study because the available phage-lambda library of purified MIC DNA was made using this strain. Strain 51 only differs at a few loci from strain d4-2 that was used for sequencing the MAC genome [77].

For gene silencing, we used the «feeding» method described in [78]. *Escherichia coli* HT115 [79] harboring plasmid L4440 [80], with the 567-bp *Hind*III-*Nco*I fragment of gene *PGM* inserted between two convergent T7 promoters [21], was induced at 37°C for the production of *PGM* dsRNA in WGP1X medium containing 100 µg/mL ampicillin. As a control, we induced HT115 bacteria for the production of dsRNA homologous to the *ND7* non essential gene (see plasmid description in [81]).

Paramecium tetraurelia strain 51new mt8 [24] was grown at 27°C in WGP1X inoculated with *Klebsiella pneumoniae* and supplemented with 0.8 µg/mL β-sistosterol. Following ~25 divisions, cells were washed and transferred to 4.1 L of freshly induced *E. coli* HT115. Cells were allowed to grow for 8 vegetative divisions, then starved to trigger autogamy. The progression of autogamy was monitored by DAPI staining (Figure S2A) and the viability of sexual progeny was tested to evaluate the efficiency of *PGM*-silencing (Figure S2B).

Cell lysis and purification of developing MAC DNA. Following prolonged starvation to favor the degradation of old MAC fragments (day 4 of autogamy), all cultures were filtered through eight layers of sterile gauze. Cells were collected by low-speed centrifugation (285× g for 1 min) and washed twice in 10 mM Tris-HCl pH 7.4. Particular care was taken to eliminate contaminating bacterial biofilms by letting them settle to the bottom of the tubes and removing them with a Pasteur pipette prior to all washing centrifugation steps. The final pellet was diluted 5-fold by addition of lysis buffer (0.25 M sucrose, 10 mM $MgCl_2$, 10 mM Tris pH 6.8, 0.2% Nonidet P-40) and processed as described in [82]. All steps were performed at 4°C. Briefly, cells (1 mL) were lysed with 100 strokes of a Potter-Elvehjem homogenizer and washing buffer (0.25 M sucrose, 10 mM $MgCl_2$, 10 mM Tris pH 7.4) was added to a final volume of 10 mL. Developing new MACs (together with cell debris, bacterial biofilms and the largest fragments of the old MAC) were collected by centrifugation at 600× g for 1 min and washed 3 times in washing buffer. To remove contaminating bacteria, the pellet was diluted in washing buffer, loaded on top of a 3-mL sucrose layer (2.1 M sucrose, 10 mM $MgCl_2$, 10 mM Tris pH 7.4) and centrifuged in a swinging rotor for 1 hr at 210,000× g. The nuclear pellet was collected and diluted 5-fold in 10 mM $MgCl_2$ 10 mM Tris pH 7.4 prior to addition of two volumes of proteinase K buffer (0.5 M EDTA pH 9, 1% N-lauryl sarcosine sodium, 1% SDS, 1 mg/mL proteinase K). Following 16-hr incubation at 55°C, genomic DNA was purified as described in [24], with three additional phenol:$CHCl_3$ extractions (1:1), one $CHCl_3$ extraction and a final ethanol precipitation [83]. Enrichment for non-excised IESs (IES^+ forms) was assayed by 1% agarose gel electrophoresis of *Pst*I-restricted DNA and Southern blot hybridization with ^{32}P-labeled Gmac probe [21], which corresponds to the MAC sequences just downstream of IES 51G4404 within the surface antigen G^{51} gene (Figure S3A). To measure the contamination with bacterial DNA, the same blot was dehybridized and probed with a ^{32}P-labelled fragment of *K. pneumoniae* 23S rDNA amplified by PCR using primers KP23S-U (5′-AGCGTTCTG-TAAGCCTGCGAAGGTG-3′) and KP23S-R (5′-TTCACCTA-CACACCAGCGTGCCTTC-3′) (Figure S3B). All radioactive

signals were scanned and quantified using a Typhoon phosphorimager (Figure S3C).

Purification of wild-type micronuclear DNA from a lambda-phage library

A lambda-phage library was provided by John Preer. This library had been made from DNA obtained after isolation of stock 51 wild type micronuclei [82] and further purified by cesium chloride density gradient centrifugation to eliminate G+C-rich DNA supposed to represent bacterial contaminants [28]. The library consisted of 70,000 recombinant phages (lambdaGEM11), expected to represent a 7× coverage of the MIC genome. We amplified the original library in 1995 and stored it at 4°C. Phage particles from 1 mL of the reamplified library (approximately 10^5 particles) were fully recovered by ultracentrifugation (42 min at 113898 g in a TLA-55 rotor; $S_{\text{lambda particle}} = 410$ according to [84]) and concentrated in ~30 µL. Given the limited amount of material (~18 pg of 40 Kb phage genomes corresponding to ~4.5 pg of inserts), the cloned DNA was amplified by PCR using primers located next to the cloning sites (LambdaL2 GGCCTAATACGACTCACTATAGG; LambdaR2 GCCATTTAGGTGACACTATAGAAGAG). Non-genomic sequences should only represent 0.6% of the total PCR-amplified DNA. As PCR inhibitors prevented direct amplification from the concentrated suspension of phage particles, 230 50 µL-PCR reactions were performed from 3 µL of a 30× dilution in SM. The Expand Long-Template PCR System (Roche) was used as recommended by the supplier with 23 amplification cycles, an annealing temperature of 60°C and 12 min for the extension time. PCR reactions were concentrated by ethanol precipitation and ~35 µg of 9 to 13 Kb PCR products were obtained after purification from 0.6% low-melting-temperature agarose gels and treatment with β-agarase (Sigma).

DNA sequencing

DNA was sequenced by a paired-end strategy using Illumina GAII and HiSeq next-generation sequencers. The shotgun fragments were ~500 bp and the paired-end reads 108 nt for DNA enriched in un-rearranged sequences (PGM DNA). The fragments were ~200 bp and the paired-end reads 101 nt for DNA prepared from the lambda-phage library. In the latter case, short reads that overlapped were merged.

Short read mapping

All Illumina short reads were mapped to the strain 51 reference genome (see below) using BWA [85] (version 0.5.8). Alignments were indexed using samtools [86] (version 0.1.11).

Strain 51 reference genome

The *P. tetraurelia* MAC genome [1] was assembled from 13× Sanger sequencing reads from different insert size librairies of strain d4-2 DNA. Strain d4-2 only differs from strain 51 at a few loci. We corrected sequencing errors in the scaffolds using Illumina deep sequencing in two stages, the first stage using the same strain d4-2 DNA sample that had been used for the original Sanger sequencing (84 million 75 nt paired-end reads), the second stage using two different samples of strain 51 MAC DNA (155 million 75 nt paired-end reads). The electronic polishing pipeline used for each stage consisted of the following steps. (i) Gap filling was achieved by assembling the Illumina reads into contigs using the Velvet [87] short read assembler (Kmer = 55 -ins_length 400 -cov_cutoff 3 -scaffolding no). The contigs were mapped to the draft assembly using BLAT [88] and locally realigned with Muscle [89]. If the contigs spanned a sequencing gap, then it was filled. (ii) The Illumina reads were mapped to the draft genome using BWA [85]. (iii) Alignments were indexed using samtools [86]. (iv) Samtools mpileup program and homemade Perl scripts were used to identify all positions covered by at least 10 reads and where at least 80% of the reads did not confirm the reference sequence. (v) The reference sequence was corrected using the list of errors. Steps (ii) through (v) were repeated a few times at the second stage of correction using the strain 51 reads, since BWA mapping has low error tolerance, and more reads could be mapped as the correction progressed. At the end of the process 442 of 861 sequencing gaps were filled, 13,758 substitutions were corrected, 929 deletions of 1–2 nt were filled and 10,339 insertions of 1–2 nt were removed, to yield the strain 51 reference genome that was used for IES identification. The strain 51 reference genome is available via ParameciumDB [90].

IES identification pipeline

MIRAA pipeline. All reads of the DNA enriched in un-rearranged sequences (PGM DNA), were mapped on the *P. tetraurelia* strain 51 reference MAC genome. Alignments indexed with samtools were analysed using custom perl scripts written with the BioPerl library (version 1.6) and the Bio::DB::Sam module (version 1.11). An IES site is characterized by an excess of ends of read alignments since reads that overlap IES junctions only map partially on the MAC genome and stop on the residual TA. These positions are considered to be IES sites if (i) the number of alignment ends is greater than 15 (10% of the average PGM DNA read coverage); (ii) if they are more than 500 bp from the ends of a scaffold, which avoids errors produced by heterogeneity in these regions; (iii) if the read coverage is lower than 300×, to avoid highly repeated sequences.

MICA pipeline. The IES detection pipeline consists of the following steps: (i) paired-end read assembly with Velvet [87] (version 1.0.18) using 3 different Kmer values (41, 45 and 55) and the parameters "-scaffolding no -max_coverage 500 -exp_cov auto -ins_length 500 -min_contig_length 100"; (ii) Only contigs with average G+C content less than 0.5 are retained; (iii) repeats are masked with RepeatMasker; (iv) masked contigs are aligned on the reference MAC genome with BLAT (version 34); (v) gaps are realigned locally with Muscle (version 3.7) and a custom Perl script is used to adjust the ends of the alignment. If the alignment is bound by TA dinucleotides, the insert in the contig is considered to be an IES. This pipeline was used to find IESs in the following sets of reads:

1. All the PGM reads after removal of known contaminants (bacteria, rDNA, mitochondrial DNA).
2. All pairs of reads in which at least one read does not align with the MAC reference genome, in order to enrich in MIC reads.
3. All PGM reads after removal of reads that correspond to the potential MAC IES junctions identified by the MIRAA pipeline.
4. Finally, all PGM reads after removal of reads that correspond to a MAC junction identified by the MICA pipeline using the above data sets.

The IES identification pipelines, datasets and overlap between IESs and potential IES sites are summarized in Figure S1. The statistics for each of the assemblies are provided in Table S1.

IES conservation

Determination of IESs that are conserved in genes duplicated by a WGD event involved identification of the position of the IES

with respect to the beginning of the alignment, either using a protein alignment of ohnologs, back translated into nucleotide sequence, or using nucleotide alignment of the 2 genes. In both cases, the alignments were carried out using Muscle [89] (version 3.7). If the relative positions of the IES is the same within a 2 nt tolerance, then the IESs are considered to be conserved.

Measurement of protein divergence

A phylogenetic tree was computed by concatenation of the alignment of 1350 protein families corresponding to quartets of ohnologs preserved after both the intermediate and recent WGD events. All gap-containing sites were excluded from the alignment, which is therefore robust with respect to possible annotation errors. The tree was constructed using BioNJ [91] with Poisson correction for multiple substitutions. The average length of the 2 branches between the intermediate and recent WGDs is 0.085 substitutions/site. The average length of the 4 branches between the recent WGD and the present is 0.0825 substitutions per site. Assuming a constant substitution rate, we can infer that the time between the intermediate and recent WGD events and between the recent WGD and the present are equivalent, although we cannot date the events since we do not know the substitution rate in *Paramecium*.

Availability of data

The MAC reference genome used for this study (strain 51) and the genome-wide set of IESs are available at http://paramecium.cgm.cnrs-gif.fr/download/. The IESs have also been integrated into ParameciumDB BioMart complex query interface and the ParameciumDB Genome Browser [90]. The short read datasets have been deposited at the European Nucleotide Archive (Accession numbers ERA137444 and ERA137420).

Validation by PCR of individual IES or IES insertion sites

Oligonucleotides were designed to flank the IES insertion site at a distance of 150–200 nt to allow detection of amplification products with or without an IES. All PCR amplifications were performed with an Eppendorf personal mastercycler. Standard PCR amplifications were performed with 1 unit of DyNazyme II with reagent concentrations according to instructions provided by Finnzyme (dNTP: 200 µM each, primers: 0.5 µM each) with 50 ng of template DNA. The program used is 2 min at 95°C, 10 cycles of 45 sec at 95°C, 45 sec at annealing temperature, and 1 min at 72°C, 15 cycles of 20 sec at 95°C, 20 sec at annealing temperature and 1 min at 72°C, followed by a final incubation for 3 min at 72°C. Amplified products were analyzed on 3% Nusieve (Lonza) in TBE 1×. Long and AT-rich PCR amplifications were performed with 1 unit of Phusion (Finnzymes) using the following concentrations of reagents (dATP and dTTP: 400 µM each, dCTP and dGTP: 200 µM each, primers: 0.5 µM each) with 50 ng of template DNA. The program used was 1 min at 98°C, 25 cycles of 10 sec at 98°C, 30 sec at annealing temperature and 5 min at 72°C, followed by a final incubation of 2 min at 72°C. Amplified products were analyzed on 1% UltraPure agarose (Invitrogen) in TAE 1×. The template DNA for the amplification reactions was an aliquot of PGM DNA enriched in un-rearranged sequences, prepared as described above.

Transposon identification

Isolation of inserts from the MIC lambda-phage library [28] was carried out as previously described [31]. Phage inserts and long-range PCR products obtained by amplification of total DNA from vegetative cells were isolated and subjected to Sanger sequencing as in [34]. The lambda-phage inserts and the cloned long-range PCR products used to characterize the *Sardine* and *Thon* transposons have been deposited in the EMBL Nucleotide Sequence Database with accession numbers HE774468–HE774475.

Several IESs with homology to the PFAM DDE_3 domain were used to find other IESs sharing nucleotide identity, leading to a set of 28 IESs that were aligned with Muscle [89] to identify 2 *Anchois* transposons. In a second step, the alignment was refined and manually adjusted in order to reconstruct the *AnchoisA* and *AnchoisB* transposons. These second step alignments were built using IESs along with some PGM contigs that correspond to germline-restricted, imprecisely eliminated regions of the genome containing *Anchois* copies (Text S1).

Data analysis

Statistical analyses and graphics were performed in the R environment for statistical computing [92] using standard packages, as well as the ape package [93] for phylogenetic analysis. Sequence logos were generated using weblogo software [94].

Supporting Information

Figure S1 IES identification. A. Schematic representation of the MIRAA pipeline for identification of IES sites by read mapping. B. Schematic representation of the MICA pipeline for identification of IESs by comparison of contigs with the reference genome assembly. C. PGM DNA datasets which were used with the MICA pipeline to identify the genome-wide set of IESs. As explained in Materials and Methods, the 4 datasets are (i) all PGM reads after filtering known contaminants, (ii) all filtered reads with at least one member of the pair that does not match the MAC reference genome, (iii) all filtered reads after removal of the read pairs with a perfect match to a MAC IES juction identified with the MIRAA pipeline and (iv) all filtered reads after removal of the read pairs with a perfect match to a MAC IES junction identified with MICA and the first 3 datasets. D.Venn diagram showing that 96% (n = 43,220) of the IESs identified with MICA correspond to IES insertion sites identified by MIRAA. The MICA pipeline was also used to identify IESs in the phage-lambda inserts: the sequence reads were assembled into 3 sets of contigs with Velvet, using 3 different kmer values (kmer = 45, 51 or 55).
(PDF)

Figure S2 Autogamy time-course of *P. tetraurelia* 51 mt8 submitted to RNAi against *PiggyMac*. A. Cells were transferred at day 0 into 4.1 L of freshly induced feeding bacteria producing dsRNA homologous to a 567-bp region of the *PGM* gene and incubated at 27°C. The progression of autogamy was monitored everyday (D1: day 1, D2: day 2, D3: day 3, D4: day 4) by DAPI staining of cells. V: vegetative cells, F: cells with fragmented old MAC and no clearly visible new developing MACs, A: cells harboring two developing new MACs, C: post-autogamous cells with one new MAC surrounded with fragments of the old MAC. B. Survival of post-autogamous progeny. At day 4, 30 autogamous cells were transferred individually to standard growth medium containing *K. pneumoniae* and incubated at 27°C to follow the resumption of vegetative growth. Survival of the progeny of autogamous cells obtained in standard (Kp) or in control RNAi medium (ND7) was also tested. Wt: normally-growing progeny, sick: slowly-growing cells, often with abnormal swimming behavior.
(PDF)

Figure S3 Purification of IES-enriched genomic DNA from PGM-silenced cells. Autogamous cells were collected at day 4 and

genomic DNA was extracted through several cell fractionation steps. Lys.1 and lys.2: independent samples of cells were lysed directly in proteinase K buffer; low sp.: DNA extracted from low speed pellets (600× g for 1 min followed by washing); suc.: DNA extracted from nuclear pellets obtained following centrifugation through a 2.1 M sucrose layer. Each DNA sample was digested by PstI and the digestion fragments were separated on a 1% agarose gel. A. Southern blot hybridization with the Gmac probe (shown as a grey box on the diagram). The position of size markers is shown on the left. IES^- and IES^+ bands were quantified separately. B. Southern blot hybridization with the *K. pneumoniae* 23S rDNA probe. Size markers are shown on the right. All rDNA bands were quantified together. C. Quantification of radioactive signals from the blots shown in A and B. The fraction of IES^+ form was normalized relative to the sum of IES^- and IES^+ signals (black histograms). Bacterial rDNA was normalized relative to the sum of IES^- and IES^+ signals (grey histograms).
(PDF)

Figure S4 IES distribution on the 8 largest MAC chromosomes. The 8 largest, telomere-capped scaffolds (~750 Kb to ~980 Kb in size) were normalized to length 1.0 and some were flipped so that the highest IES density is to the right. The curves represent histograms of IES position on each scaffold after Gaussian smoothing using the R "density" function [92]. IES distribution was evaluated using a Kolmogorov-Smirnov test of the null hypothesis that IESs are uniformly distributed on the scaffold. For the 8 largest scaffolds, the null hypothesis was strongly rejected ($p<10^{-8}$). The same statistical test was carried out for gene distribution on these chromosomes, and the null hypothesis was not rejected, consistent with a uniform distribution of genes on the chromosomes.
(PDF)

Figure S5 *Sardine* and *Thon* Tc1/mariner family transposons. From top to bottom: 1) *Sardine* transposon consensus sequence obtained by alignment of the lambda-phage and PCR copies (the latter were amplified from total DNA of vegetative cells using primers located within the *Sardine* TIRs), showing the presence of palindromic TIRs and 4 putative ORFs, including a DDE transposase and a tyrosine recombinase; 2) lambda-phage with the 51G flank that led to discovery of the *Sardine* element (the region of *de novo* telomere addition at the end of the MAC chromosome, following developmental breakage of the MIC chromosome, is indicated); 3) lambda-phage with the S5 copy of *Sardine*; 4) lambda-phage with the S6 copy of *Sardine*, containing an insertion of a different Tc1/mariner transposon, *Thon*, which has the same general organization as the *Sardine* element; 5) lambda-phage with the S7 copy of *Sardine*; 6) lambda-phage with the S8 copy; 7) PCR products (S46 and S103 copies) with nearly intact ORFs; 8) PCR products (S14 and S106 copies) with nearly intact ORFs. The sequences of the 5 lambda-phages and 4 PCR products have been deposited in the EMBL/GenBANK/DDBJ public nucleotide database with EMBL-Bank accession numbers HE774468–HE774475.
(PDF)

Figure S6 IES size distribution. The histograms represent A) IESs inserted in coding sequences. B) IESs inserted in non-coding sequences. IESs larger than 150 nt are not displayed. The fact that very similar periodic distributions are observed for IESs in both coding and intergenic regions is consistent with the hypothesis that the periodic size constraint is related to the IES excision mechanism. Indeed, IES retention in the MAC could be deleterious either by affecting ORFs (IESs in protein coding sequences) or by affecting regulatory signals (IESs in non-coding sequences).
(PDF)

Figure S7 IES evolution evaluated with quartet IES groups. A) schematic representation of the observable quartet IES groups, arranged from top to bottom according to the number of IESs that are conserved and from left to right, according to the most recent period in which the ancestral IES could have been acquired. B) Schematic representation of the parameters of a statistical model developed to test hypotheses about IES evolution (cf. Text S1). The three time periods delimited by the 2 WGD events and the present time are designated, from the oldest to the most recent, g_3, g_2 and g_1. The parameters ρ_3, ρ_2 and ρ_1 are the fraction of IESs that were acquired in each of these time period and the parameters of the form $\delta_{a,b}$ are the survival rates for an IES acquired in period g_a during the period g_b. The equations of the model express the observable IES counts as a function of these parameters.
(PDF)

Table S1 Assembly statistics.
(PDF)

Table S2 Molecular validation of some predicted IESs and IES insertion sites.
(PDF)

Table S3 Validation of the genome-wide set of IESs using previously characterized IESs.
(PDF)

Table S4 IESs with homology to *Anchois* transposons.
(PDF)

Table S5 Deficit of 3n IESs in coding sequences, for each peak of the 10 bp periodic size distribution.
(PDF)

Text S1 Transposon sequences. A). The sequences of *Sardine*, *Thon* and *Anchois* transposons reconstituted from manually adjusted multiple alignments of the different decayed copies, cloned from the lambda phage library of MIC DNA (*Sardine*, *Thon*) or found in the PGM DNA assembly (*Anchois*). The sequences of the *Thon* transposon are those of the only known copy, so that ORF annotation (based on homology with the *Sardine* element) is preliminary; the *Thon* ORF1 sequence apparently contains a frameshift. Predicted introns have been removed from the ORF sequences. B) Annotated comparison of *AnchoisA* and *AnchoisB*, showing the position and orientation of the ORFs, with a potential intron in the DDE transposase ORF. C). Manually adjusted alignment used to reconstitute the AnchoisA copy. See Text S4 for the IESs used in the reconstitution. D). Manually adjusted alignment used to reconstitute the *AnchoisB* copy. See Text S4 for the IESs used in the reconstitution. E) IESs used to obtain the final *AnchoisA* and *AnchoisB* consensus sequences based on the manually adjusted alignments in C) and D).
(PDF)

Text S2 Alignment of homologous IESs inserted at non-homologous genomic sites. The IESs of each cluster of homologous IESs (cf. Table 3 and its legend) and 200 bp of 3′ and 5′ flanking sequences were aligned using Muscle [89]. The IESs are in uppercase type and the flanking sequences are in lowercase type. For cluster5, consisting of IESs homologous to a solo TIR of the *Thon* transposon, the consensus sequence and the *Thon* TIR are included in the alignment and the palindromic repeats are highlighted.
(PDF)

Text S3 PCR approach to validate IESs with homology to *Thon* solo TIRs.
(PDF)

Text S4 A maximum likelihood framework for testing hypotheses about IES evolution.
(PDF)

Acknowledgments

The authors thank Laura Katz, Feng Gao, and Deepankar Pratap Singh for permission to cite their unpublished data and Jean Cohen, Emeline Dubois, and Julien Bischerour for critical reading of the manuscript. The project was carried out in the framework of the CNRS–supported European Research Group "Paramecium Genome Dynamics and Evolution" and the European Science Foundation COST network BM1102 "Ciliates as model systems to study genome evolution, mechanisms of non-Mendelian inheritance, and their roles in environmental adaptation."

Author Contributions

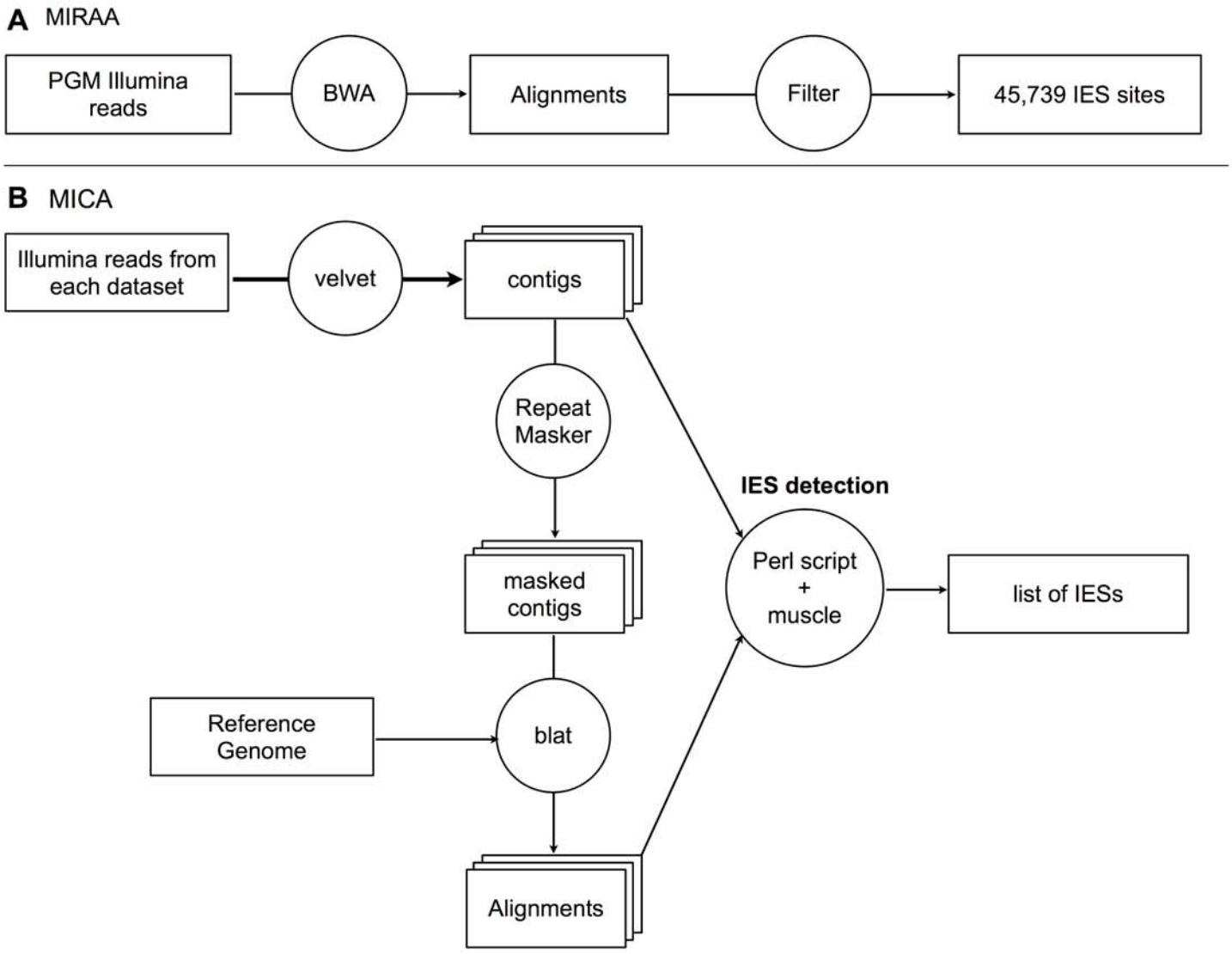
Conceived and designed the experiments: MB LD SD EM SM LS. Performed the experiments: MB CB SD CDW NM SM AM MN OG ALM MP EM. Analyzed the data: OA CDW LD LS. Wrote the paper: OA MB LD SD BEL EM SM LS. Mathematical model: BEL. DNA sequencing: J-MA KL JP PW.

References

1. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444: 171–178. doi:nature05230.
2. Chalker DL, Yao M-C (2011) DNA elimination in ciliates: transposon domestication and genome surveillance. Annu Rev Genet 45: 227–246. doi:10.1146/annurev-genet-110410-132432.
3. Coyne RS, Lhuillier-Akakpo M, Duharcourt S (2012) RNA-guided DNA rearrangements in ciliates: is the best genome defense a good offense? Biol Cell Accepted manuscript online. doi:10.1111/boc.201100057.
4. Schoeberl UE, Mochizuki K (2011) Keeping the soma free of transposons: programmed DNA elimination in ciliates. J Biol Chem 286: 37045–37052. doi:10.1074/jbc.R111.276964.
5. Bétermier M (2004) Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate Paramecium. Res Microbiol 155: 399–408.
6. Ruiz F, Krzywicka A, Klotz C, Keller A, Cohen J, et al. (2000) The SM19 gene, required for duplication of basal bodies in Paramecium, encodes a novel tubulin, eta-tubulin. Curr Biol 10: 1451–1454.
7. Haynes WJ, Ling KY, Preston RR, Saimi Y, Kung C (2000) The cloning and molecular analysis of pawn-B in Paramecium tetraurelia. Genetics 155: 1105–1117.
8. Mayer KM, Mikami K, Forney JD (1998) A mutation in Paramecium tetraurelia reveals functional and structural features of developmentally excised DNA elements. Genetics 148: 139–149.
9. Mayer KM, Forney JD (1999) A mutation in the flanking 5′-TA-3′ dinucleotide prevents excision of an internal eliminated sequence from the Paramecium tetraurelia genome. Genetics 151: 597–604.
10. Matsuda A, Forney JD (2005) Analysis of Paramecium tetraurelia A-51 surface antigen gene mutants reveals positive-feedback mechanisms for maintenance of expression and temperature-induced activation. Eukaryotic Cell 4: 1613–1619. doi:10.1128/EC.4.10.1613-1619.2005.
11. Yao MC, Choi J, Yokoyama S, Austerberry CF, Yao CH (1984) DNA elimination in Tetrahymena: a developmental process involving extensive breakage and rejoining of DNA at defined sites. Cell 36: 433–440.
12. Saveliev SV, Cox MM (2001) Product analysis illuminates the final steps of IES deletion in Tetrahymena thermophila. EMBO J 20: 3251–3261. doi:10.1093/emboj/20.12.3251.
13. Fillingham JS, Thing TA, Vythilingum N, Keuroghlian A, Bruno D, et al. (2004) A non-long terminal repeat retrotransposon family is restricted to the germ line micronucleus of the ciliated protozoan Tetrahymena thermophila. Eukaryotic Cell 3: 157–169.
14. Wuitschick JD, Gershan JA, Lochowicz AJ, Li S, Karrer KM (2002) A novel family of mobile genetic elements is limited to the germline genome in Tetrahymena thermophila. Nucleic Acids Res 30: 2524–2537.
15. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, et al. (2006) Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol 4: e286. doi:10.1371/journal.pbio.0040286.
16. Yao M-C, Chao J-L (2005) RNA-guided DNA deletion in Tetrahymena: an RNAi-based mechanism for programmed genome rearrangements. Annu Rev Genet 39: 537–559. doi:10.1146/annurev.genet.39.073003.095906.
17. Fass JN, Joshi NA, Couvillion MT, Bowen J, Gorovsky MA, et al. (2011) Genome-Scale Analysis of Programmed DNA Elimination Sites in Tetrahymena thermophila. G3 1: 515–522. doi:10.1534/g3.111.000927.
18. Klobutcher LA, Herrick G (1995) Consensus inverted terminal repeat sequence of Paramecium IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons. Nucleic Acids Res 23: 2006–2013.
19. Klobutcher LA, Herrick G (1997) Developmental genome reorganization in ciliated protozoa: the transposon link. Prog Nucleic Acid Res Mol Biol 56: 1–62.
20. Plasterk RH, Izsvák Z, Ivics Z (1999) Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet 15: 326–332.
21. Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, et al. (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev 23: 2478–2483. doi:10.1101/gad.547309.
22. Cheng C-Y, Vogt A, Mochizuki K, Yao M-C (2010) A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in Tetrahymena thermophila. Mol Biol Cell 21: 1753–1762. doi:10.1091/mbc.E09-12-1079.
23. Parfrey LW, Lahr DJG, Knoll AH, Katz LA (2011) Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci USA 108: 13624–13629. doi:10.1073/pnas.1110633108.
24. Gratias A, Bétermier M (2003) Processing of double-strand breaks is involved in the precise excision of paramecium internal eliminated sequences. Mol Cell Biol 23: 7152–7162.
25. Mitra R, Fain-Thornton J, Craig NL (2008) piggyBac can bypass DNA synthesis during cut and paste transposition. EMBO J 27: 1097–1109. doi:10.1038/emboj.2008.41.
26. Kapusta A, Matsuda A, Marmignon A, Ku M, Silve A, et al. (2011) Highly precise and developmentally programmed genome assembly in Paramecium requires ligase IV-dependent end joining. PLoS Genet 7: e1002049. doi:10.1371/journal.pgen.1002049.
27. Preer LB, Hamilton G, Preer JR Jr (1992) Micronuclear DNA from Paramecium tetraurelia: serotype 51 A gene has internally eliminated sequences. J Protozool 39: 678–682.
28. Steele CJ, Barkocy-Gallagher GA, Preer LB, Preer JR Jr (1994) Developmentally excised sequences in micronuclear DNA of Paramecium. Proc Natl Acad Sci USA 91: 2255–2259.
29. Duret L, Cohen J, Jubin C, Dessen P, Goût J-F, et al. (2008) Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia: a somatic view of the germline. Genome Res 18: 585–596. doi:gr.074534.107.
30. Arnaiz O, Sperling L (2010) ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia. Nucleic Acids Res. Available:http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/20952411. Accessed 14 December 2010.
31. Duharcourt S, Keller AM, Meyer E (1998) Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in Paramecium tetraurelia. Mol Cell Biol 18: 7075–7085.
32. Doak TG, Doerder FP, Jahn CL, Herrick G (1994) A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc Natl Acad Sci USA 91: 942–946.
33. Jacobs ME, Sánchez-Blanco A, Katz LA, Klobutcher LA (2003) Tec3, a new developmentally eliminated DNA element in Euplotes crassus. Eukaryotic Cell 2: 103–114.
34. Le Mouël A, Butler A, Caron F, Meyer E (2003) Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in paramecia. Eukaryotic Cell 2: 1076–1090.
35. Jaillon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, et al. (2008) Translational control of intron splicing in eukaryotes. Nature 451: 359–362. doi:nature06495.
36. DuBois ML, Prescott DM (1997) Volatility of internal eliminated segments in germ line genes of hypotrichous ciliates. Mol Cell Biol 17: 326–337.
37. Dubrana K, Le Mouël A, Amar L (1997) Deletion endpoint allele-specificity in the developmentally regulated elimination of an internal sequence (IES) in Paramecium. Nucleic Acids Res 25: 2448–2454.
38. Gout J-F, Kahn D, Duret L (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet 6: e1000944. doi:10.1371/journal.pgen.1000944.
39. Arnaiz O, Gout J-F, Betermier M, Bouhouche K, Cohen J, et al. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. BMC Genomics 11: 547. doi:10.1186/1471-2164-11-547.
40. Duharcourt S, Butler A, Meyer E (1995) Epigenetic self-regulation of developmental excision of an internal eliminated sequence on Paramecium tetraurelia. Genes Dev 9: 2065–2077.
41. Meyer E, Keller AM (1996) A Mendelian mutation affecting mating-type determination also affects developmental genomic rearrangements in Paramecium tetraurelia. Genetics 143: 191–202.

42. Duharcourt S, Lepère G, Meyer E (2009) Developmental genome rearrangements in ciliates: a natural genomic subtraction mediated by non-coding transcripts. Trends Genet 25: 344–350. doi:10.1016/j.tig.2009.05.007.
43. Lepère G, Nowacki M, Serrano V, Gout J-F, Guglielmi G, et al. (2009) Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia. Nucleic Acids Res 37: 903–915. doi:10.1093/nar/gkn1018.
44. Lepère G, Bétermier M, Meyer E, Duharcourt S (2008) Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in Paramecium tetraurelia. Genes Dev 22: 1501–1512. doi:10.1101/gad.473008.
45. Nowacki M, Zagorski-Ostoja W, Meyer E (2005) Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in Paramecium tetraurelia. Curr Biol 15: 1616–1628. doi:10.1016/j.cub.2005.07.033.
46. Epstein LM, Forney JD (1984) Mendelian and non-mendelian mutations affecting surface antigen expression in Paramecium tetraurelia. Mol Cell Biol 4: 1583–1590.
47. Meyer E (1992) Induction of specific macronuclear developmental mutations by microinjection of a cloned telomeric gene in Paramecium primaurelia. Genes Dev 6: 211–222.
48. Sperling L (2011) Remembrance of things past retrieved from the Paramecium genome. Res Microbiol 162: 587–597. doi:10.1016/j.resmic.2011.02.012.
49. Schleif R (1992) DNA Looping. Annual Review of Biochemistry 61: 199–223. doi:10.1146/annurev.bi.61.070192.001215.
50. Lane D, Cavaillé J, Chandler M (1994) Induction of the SOS response by IS1 transposase. J Mol Biol 242: 339–350. doi:10.1006/jmbi.1994.1585.
51. Goryshin IYu, Kil YV, Reznikoff WS (1994) DNA length, bending, and twisting constraints on IS50 transposition. Proc Natl Acad Sci USA 91: 10834–10838.
52. Müller J, Oehler S, Müller-Hill B (1996) Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. J Mol Biol 257: 21–29. doi:10.1006/jmbi.1996.0143.
53. Bellomy GR, Mossing MC, Record MT Jr (1988) Physical properties of DNA in vivo as probed by the length dependence of the lac operator looping process. Biochemistry 27: 3900–3906.
54. Bond LM, Peters JP, Becker NA, Kahn JD, Maher LJ (2010) Gene repression by minimal lac loops in vivo. Nucleic Acids Research 38: 8072–8082. doi:10.1093/nar/gkq755.
55. Lee DH, Schleif RF (1989) In vivo DNA loops in araCBAD: size limits and helical repeat. Proceedings of the National Academy of Sciences 86: 476–480.
56. Haykinson MJ, Johnson RC (1993) DNA looping and the helical repeat in vitro and in vivo: effect of HU protein and enhancer location on Hin invertasome assembly. EMBO J 12: 2503–2512.
57. Bétermier M, Duharcourt S, Seitz H, Meyer E (2000) Timing of developmentally programmed excision and circularization of Paramecium internal eliminated sequences. Mol Cell Biol 20: 1553–1561.
58. Gratias A, Lepère G, Garnier O, Rosa S, Duharcourt S, et al. (2008) Developmentally programmed DNA splicing in Paramecium reveals short-distance crosstalk between DNA cleavage sites. Nucleic Acids Res 36: 3244–3251. doi:10.1093/nar/gkn154.
59. Chandler M, Mahillon J (2002) Insertion Sequences Revisited. Mobile DNA II. Washington, D.C.: ASM Press. pp. 305–366.
60. Paull TT, Haykinson MJ, Johnson RC (1993) The nonspecific DNA-binding and -bending proteins HMG1 and HMG2 promote the assembly of complex nucleoprotein structures. Genes Dev 7: 1521–1534.
61. Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, et al. (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res 19: 2221–2230. doi:10.1101/gr.083899.108.
62. Garfinkel DJ, Nyswaner KM, Stefanisko KM, Chang C, Moore SP (2005) Ty1 copy number dynamics in Saccharomyces. Genetics 169: 1845–1857. doi:10.1534/genetics.104.037317.
63. Lynch M (2006) The origins of eukaryotic gene structure. Mol Biol Evol 23: 450–468. doi:10.1093/molbev/msj050.
64. Werren JH (2011) Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc Natl Acad Sci USA 108 Suppl 2: 10863–10870. doi:10.1073/pnas.1102343108.
65. Bourc'his D, Voinnet O (2010) A small-RNA perspective on gametogenesis, fertilization, and early zygotic development. Science 330: 617–622. doi:10.1126/science.1194776.
66. Malone CD, Hannon GJ (2009) Molecular evolution of piRNA and transposon control pathways in Drosophila. Cold Spring Harb Symp Quant Biol 74: 225–234. doi:10.1101/sqb.2009.74.052.
67. Baulcombe D (2004) RNA silencing in plants. Nature 431: 356–363. doi:10.1038/nature02874.
68. Prescott DM, Prescott JD, Prescott RM (2002) Coding properties of macronuclear DNA molecules in Sterkiella nova (Oxytricha nova). Protist 153: 71–77.
69. Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, et al. (2008) RNA-mediated epigenetic programming of a genome-rearrangement pathway. Nature 451: 153–158.
70. Prescott DM (1999) The evolutionary scrambling and developmental unscrambling of germline genes in hypotrichous ciliates. Nucleic Acids Res 27: 1243–1250.
71. Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG, et al. (2009) A functional role for transposases in a large eukaryotic genome. Science 324: 935–938. doi:10.1126/science.1170023.
72. Jahn CL, Klobutcher LA (2002) Genome remodeling in ciliated protozoa. Annu Rev Microbiol 56: 489–520. doi:10.1146/annurev.micro.56.012302.160916.
73. Jahn CL, Doktor SZ, Frels JS, Jaraczewski JW, Krikau MF (1993) Structures of the Euplotes crassus Tec1 and Tec2 elements: identification of putative transposase coding regions. Gene 133: 71–78.
74. Jaraczewski JW, Jahn CL (1993) Elimination of Tec elements involves a novel excision process. Genes Dev 7: 95–105.
75. Klobutcher LA, Turner LR, LaPlante J (1993) Circular forms of developmentally excised DNA in Euplotes crassus have a heteroduplex junction. Genes Dev 7: 84–94.
76. Lambowitz AM, Zimmerly S (2011) Group II Introns: Mobile Ribozymes that Invade DNA. Cold Spring Harbor Perspectives in Biology 3. Available:http://cshperspectives.cshlp.org/content/3/8/a003616.abstract.
77. Sonneborn TM (1974) Paramecium aurelia. Handbook of Genetics. R. King. New York: Plenum Press, Vol. 11. pp. 469–594.
78. Galvani A, Sperling L (2002) RNA interference by feeding in Paramecium. Trends Genet 18: 11–12.
79. Timmons L, Court DL, Fire A (2001) Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. Gene 263: 103–112.
80. Timmons L, Fire A (1998) Specific interference by ingested dsRNA. Nature 395: 854. doi:10.1038/27579.
81. Garnier O, Serrano V, Duharcourt S, Meyer E (2004) RNA-mediated programming of developmental genome rearrangements in Paramecium tetraurelia. Mol Cell Biol 24: 7370–7379. doi:10.1128/MCB.24.17.7370-7379.2004.
82. Preer LB, Hamilton G, Preer JR Jr (1992) Micronuclear DNA from Paramecium tetraurelia: serotype 51 A gene has internally eliminated sequences. J Protozool 39: 678–682.
83. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning: A Laboratory Manual. 2nd ed. Cold Spring Harbor Laboratory Pr. 1659 p.
84. Weigle J (1966) Assembly of phage lambda in vitro. Proc Natl Acad Sci USA 55: 1462–1466.
85. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760. doi:10.1093/bioinformatics/btp324.
86. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. doi:10.1093/bioinformatics/btp352.
87. Zerbino D, Birney E (2008) Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. Genome Res: gr.074492.107. doi:10.1101/gr.074492.107.
88. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12: 656–664. doi:10.1101/gr.229202.
89. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797. doi:10.1093/nar/gkh340.
90. Arnaiz O, Sperling L (2011) ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia. Nucleic Acids Res 39: D632–636. doi:10.1093/nar/gkq918.
91. Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14: 685–695.
92. R Development Core Team (2011) R: A language and environment for statistical computing. Available:http://www.R-project.org.
93. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289–290.
94. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190. doi:10.1101/gr.849004.
95. Catania F, Wurmser F, Potekhin AA, Przybos E, Lynch M (2009) Genetic diversity in the Paramecium aurelia species complex. Mol Biol Evol 26: 421–431. doi:10.1093/molbev/msn266.

Figure S1



C

Dataset	No. of reads	No. of IESs
All PGM reads	126 M	44,207
Reads at least one no match	32 M	44,211
Reads without MAC junc MIRAA	121 M	44,269
Remove all MAC junc	121 M	43,286
Merge	**-**	**44,928**

D

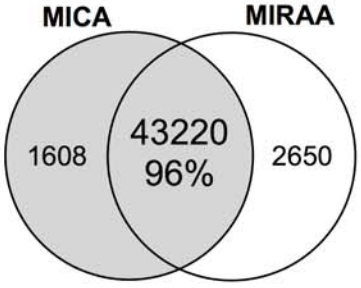


Figure S2

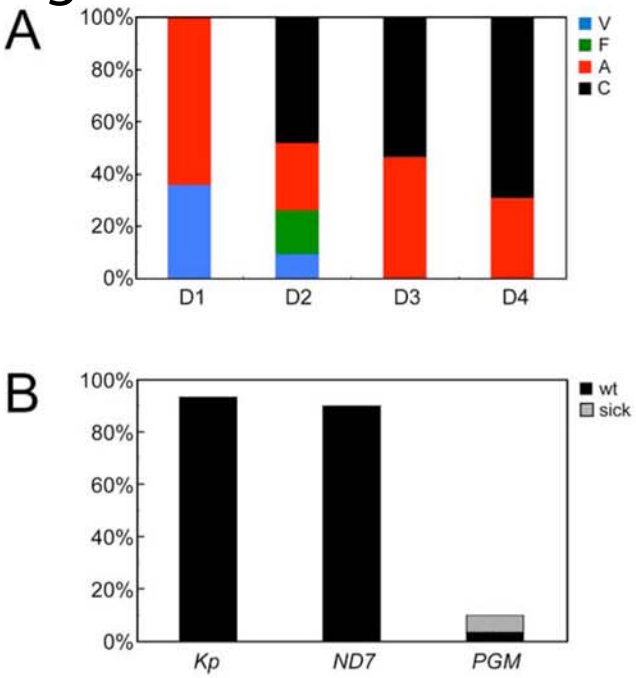


Figure S3

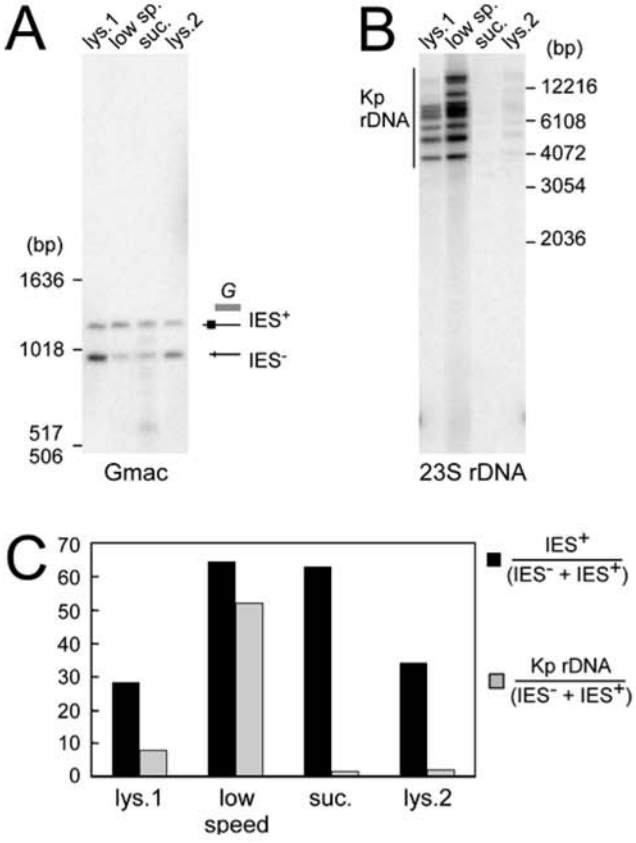


Figure S4

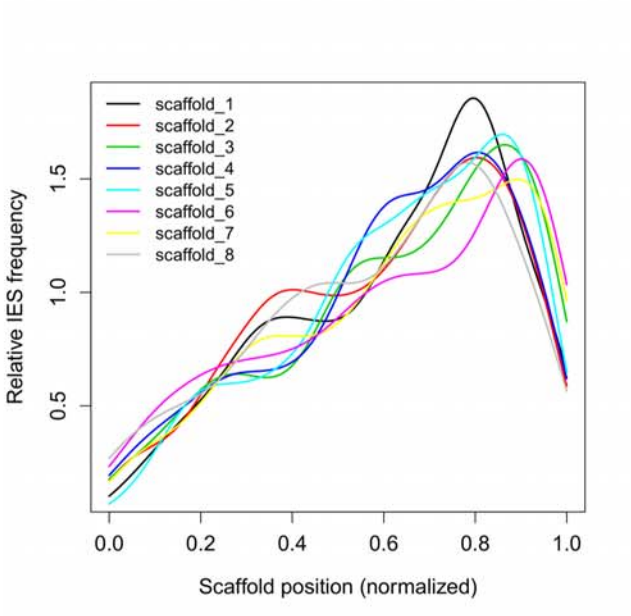


Figure S5

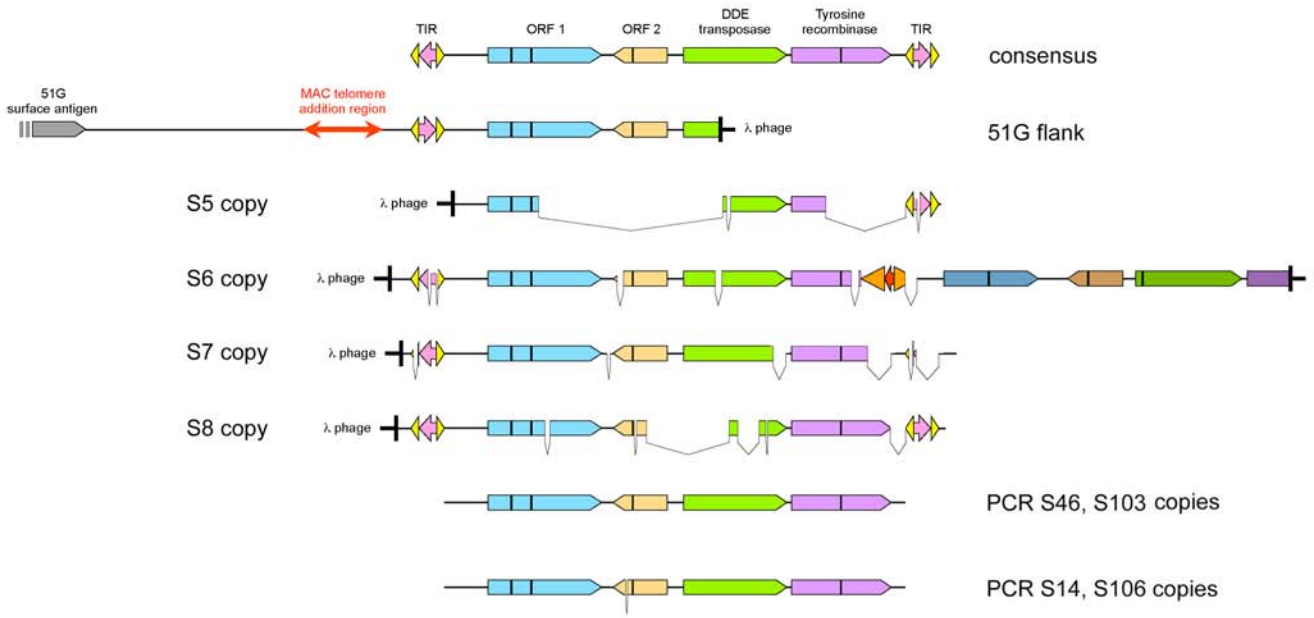


Figure S6

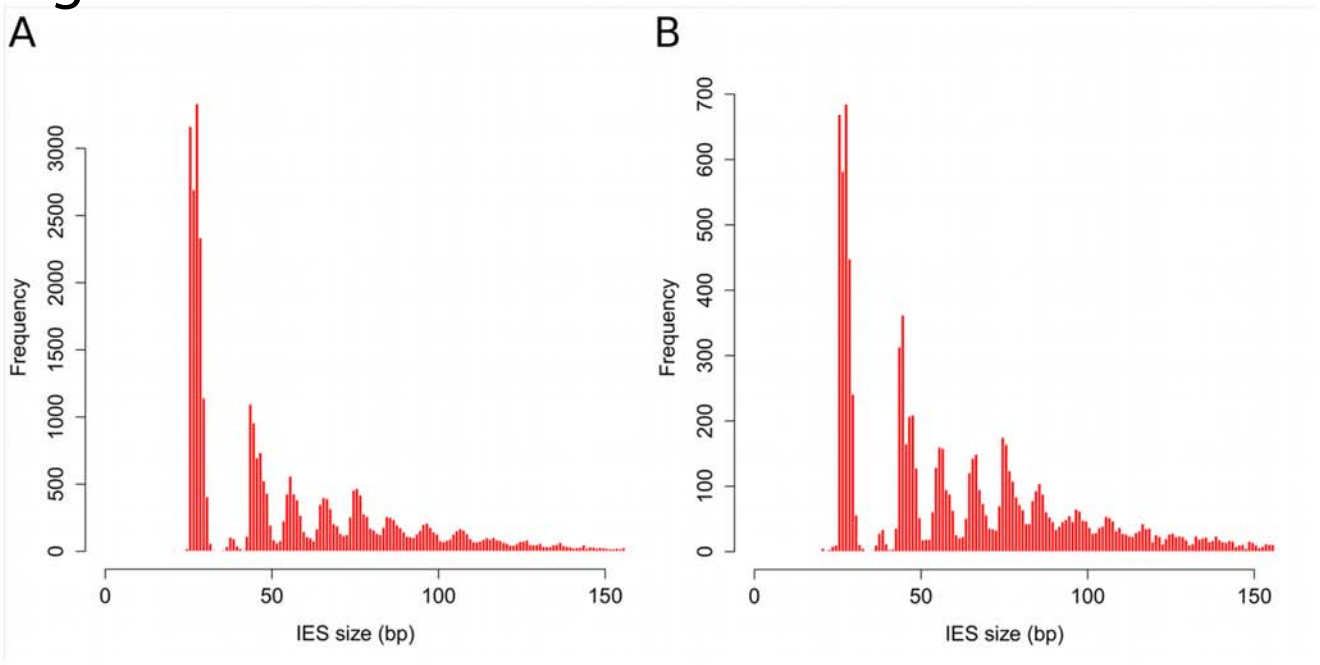
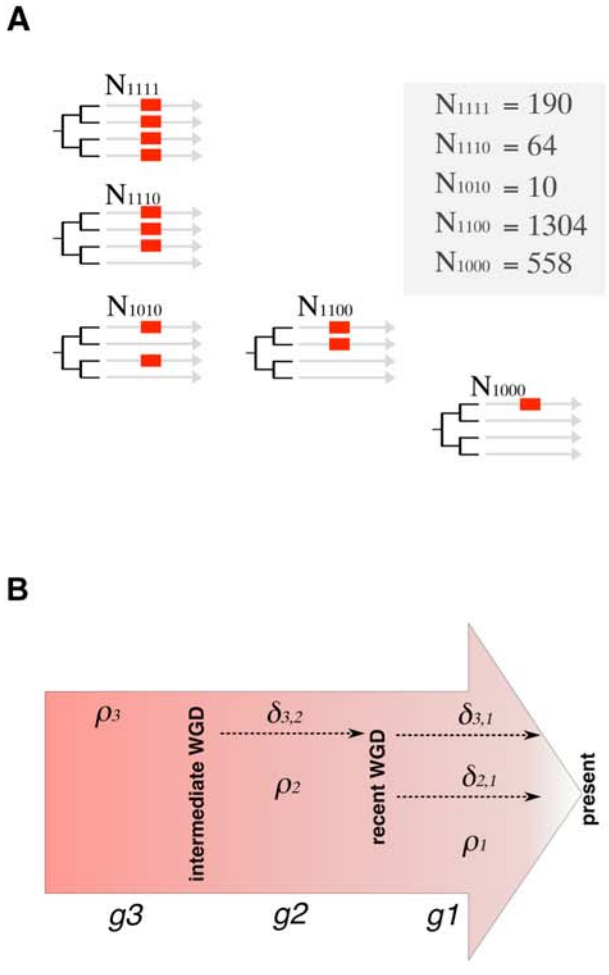


Figure S7



Description	GenBank Accession N°	Scaffold	Position	% ID	known IES length (bp)	IES length (this study) (bp)	COMMENT
ND7_IES4_mic	gi|193480009	scaffold51_5	597595	100	29	29	
ND7_IES3_mic	gi|193480009	scaffold51_5	595021	100	26	26	
ND7_IES2_mic	gi|193480009	scaffold51_5	594919	100	76	76	
ND7_IES1_mic	gi|193480009	scaffold51_5	594675	100	28	28	
IESsm19-576_mic	GB	scaffold51_8	407305	100	66	66	
IES51pwB-658_mic_alternative	gi|10716831	scaffold51_19	587496	100	146	146	
IES51pwB-427_mic	gi|10716831	scaffold51_19	587727	100	44	44	
IES51pwB-2226_mic	GB	scaffold51_19	587126	100	66	66	
IES51PAK11-991_mic	gi|3916119	scaffold51_5	724979	100	28	28	
IES51PAK11-294_mic	gi|3916119	scaffold51_5	725700	100	45	45	
IES51PAK11-1557_mic	gi|3916119	scaffold51_5	724388	100	44	44	
IES51PAK1-991_mic	gi|3894330	scaffold51_16	560295	100	27	27	
IES51PAK1-1557_mic	gi|3894330	scaffold51_16	559704	100	44	44	
IES51PAK1-1036_mic	gi|3894330	scaffold51_16	560250	100	46	46	
IES51ICL1b_mic	gi|1666902	scaffold51_134	33201	100	75	75	
IES51G6447_mic	gi|3452504	scaffold51_51	457658	100	28	28	
IES51G4404_mic	gi|3452504	scaffold51_51	455615	100	222	222	
IES51G2832_mic	gi|3452504	scaffold51_51	454043	100	229	229	
IES51G1832_mic	gi|3452504	scaffold51_51	453043	100	30	30	
IES51G1413_mic	gi|3452504	scaffold51_51	452624	100	52	52	
IES51G-11_mic	gi|3452504	scaffold51_51	451201	100	43	43	
IES51B5464_mic	gi|76589367	scaffold51_143	33517	100	28	28	
IES51B1417_mic	gi|76589367	scaffold51_143	37564	100	44	44	
IES51B-9_mic	gi|76589367	scaffold51_143	38990	100	36	36	
IES51A6649_mic	gi|78707424	scaffold51_106	288995	100	370	370	
IES51A6435_mic	gi|78707424	scaffold51_106	288781	100	28	28	
IES51A4578_mic	gi|78707424	scaffold51_106	286924	100	883	883	
IES51A4404_mic	gi|78707424	scaffold51_106	286750	100	77	77	
IES51A2591_mic	gi|78707424	scaffold51_106	284913	100	370	370	
IES51A1835_mic	gi|78707424	scaffold51_106	284157	100	28	28	
IES51A1416_mic	gi|78707424	scaffold51_106	283738	100	74	74	
IES51A-712_mic	gi|78707424	scaffold51_106	281631	100	77	77	
IES51A-4814_mic	gi|29469820	scaffold51_106	277532	100	28	28	
IES51A-10_mic	gi|78707424	scaffold51_106	282313	100	29	29	
epsilon51D_28bp_mic	gi|1752672	scaffold51_128	10690	100	28	28	
alpha51D_IES5_mic	gi|1752672	scaffold51_159	213381	100	28	28	
alpha51D_IES4_mic	gi|1752672	scaffold51_159	209504	100	44	44	
alpha51D_IES3_mic	gi|1752672	scaffold51_159	207374	100	28	28	

Dataset	Million reads	Kmer	No. of Contigs	Largest contig (nt)	Complexity (nt)	N50	MAC genome coverage
1	126	45	38714	159958	96838332	18877	96.29%
1	126	51	32891	177346	97451433	26902	96.38%
1	126	55	42969	82851	99085030	7298	97.30%
2	32	45	94491	107860	61535444	1087	49.74%
2	32	51	80141	86502	59448180	1180	46.45%
2	32	55	71678	82851	58477670	1230	44.89%
3	121	45	26721	158845	95582645	24907	95.91%
3	121	51	23091	187719	96119509	29755	95.65%
3	121	55	22059	138361	96437319	25160	96.63%
4	121	55	20728	243532	96552463	32589	96.42%

Table S1. Assembly statistics. The datasets are as follows : 1) All PGM reads; 2) All pairs of PGM reads with at least one read that does not map to the MAC reference genome; 3) All PGM reads after removal of those reads that do not match a putative MAC IES junction identified by the MIRAA pipeline; 4) All PGM reads after removal of those reads that do not match a MAC IES junction identified using the MICA pipeline and the first 9 assemblies. Paired-end assembly was carried out using the Velvet short read assembler (version 1.0.18) with the indicated Kmer value. The MAC genome coverage by the assembled contigs was determined by mapping them to the reference genome with BLAT.

Type	Scaffold	Position	MAC PCR product		MIC PCR product		IES size	Sequenced
			Expected	Observed	Expected	Observed		
MICA & MIRAA	scaffold51_1	292806	369	~375	454	~460	85	no
MICA & MIRAA	scaffold51_10	240455	355	~360	432	~450	77	no
MICA & MIRAA	scaffold51_11	141916	364	~375	573	~560	209	no
MICA & MIRAA	scaffold51_12	366211	350	~350	377	~375	27	MIC
MICA & MIRAA	scaffold51_14	132164	361	~360	408	~410	47	no
MICA & MIRAA	scaffold51_2	623719	348	~350	404	~410	56	no
MICA & MIRAA	scaffold51_3	554680	367	~360	403	~400	36	MIC
MIRAA	scaffold51_1	206570	322	~330	ND	~350	*31-2-28	MIC
MIRAA	scaffold51_11	460330	375	~400	ND	~450	*29-1-29	MIC
MIRAA	scaffold51_13	596382	347	~350	ND	~400	*28-2-30	MIC
MIRAA	scaffold51_14	210046	366	~375	ND	~800	ND	no
MIRAA	scaffold51_15	406155	369	~375	ND	~650	ND	no
MIRAA	scaffold51_16	361775	350	~350	ND	~400	*26-3-26	MIC
MIRAA	scaffold51_9	114280	342	~350	ND	~460	*57-1-46	MIC
MIRAA	scaffold51_56	454715	332	~340	ND	~3200	ND	no
MIRAA	scaffold51_47	324362	358	~360	ND	~410	*47-4-29	MIC
MIRAA	scaffold51_34	257109	394	~400	ND	~650	*106-6-148	MIC
MIRAA	scaffold51_41	194697	378	~400	ND	~475	*28-3-58	MIC
MIRAA	scaffold51_26	330923	345	~360	ND	~500	146	MIC
MIRAA	scaffold51_35	149456	358	~375	ND	none	ND	no
MIRAA	scaffold51_43	182962	366	~400	ND	none	ND	no
MIRAA	scaffold51_37	205471	343	~350	ND	none	ND	no
MIRAA	scaffold51_33	261960	349	~350	ND	none	ND	MAC
MIRAA	scaffold51_28	472294	358	~400	ND	none	ND	no
MICA	scaffold51_179	88164	333	~340	360	~375	27	no
MICA	scaffold51_184	64423	308	~325	353	~350	45	no
MICA	scaffold51_188	26049	351	~360	378	~380	27	no
MICA	scaffold51_188	29158	346	~350	390	~400	44	MIC
MICA	scaffold51_71	364423	373	~390	400	~480	27	no
MICA	scaffold51_87	125232	358	~360	385	~390	27	no

Table S2. Molecular validation of some predicted IESs and IES insertion sites. IES insertion sites predicted by one or both methods (MICA and/or MIRAA) were tested using PCR as described in Materials and Methods. For each tested MICA prediction, the MIC PCR product was of the expected size, and for those that were sequenced, the expected IES sequence was found. For insertion sites predicted only by MIRAA, PCR amplifications show unexpected results : IESs very close one to another, spaced by only one or a few nucleotides, were often found. Out of 17 tested MIRAA predictions, 8 appeared to be such tandem IESs, 1 appeared to be an 146 bp IES, in 5 cases we did not obtain a MIC amplification product, and in the remaining 3 cases, the MIC amplification product was not sequenced. Failure to dectect a MIC product does not exclude the presence of a large IES, given the PCR conditions that were used.

Legend.
* Tandem IESs with size format : IES_1_size – space_size – IES_2_size.
~ Approximate size of the corresponding band in the migration gel.
ND (Expected MIC): we could not predict any size for a MIC amplification because no *a priori* prediction was made on the presence of a MIC sequence.
none (Product MIC): no amplification other than the MAC product could be observed.
ND (IES size): the MIC product was not sequenced or its size could not be predicted.

alpha51D_IES1_mic	gi|1752672	scaffold51_159	205692	100	26	26	
51ICL1d_IES3_mic	gi|1667584	scaffold51_124	172613	100	29	29	
51ICL1d_IES2_mic	gi|1667583	scaffold51_124	173737	100	45	45	
51ICL1d_IES1_mic	gi|1667583	scaffold51_124	173840	100	26	26	
IES51pwB-658_mic	**gi|10716831**	**scaffold51_19**	**587496**	**100**	**155**	**146**	**alternative boundary form not found**
IES51A6649_29bp_mic	**gi|78707424**	**NA**	**NA**	**0**	**29**	**NA**	**nested IES**
IES51A2591_28bp_mic	**gi|78707424**	**NA**	**NA**	**0**	**28**	**NA**	**nested IES**
IES51A15885_mic	**GB**	**NA**	**NA**	**0**	**54**	**NA**	**assembled in MAC reference**
IES51pwB-383_mic	GB	scaffold51_19	588978	100	47	49	
IES51PAK11-1036_mic	gi|3916119	scaffold51_5	724934	98.1	50	50	
IES51PAK1-538_mic	gi|3894330	scaffold51_16	560748	99.5	206	206	
IES51PAK1-294_mic	gi|3894330	scaffold51_16	561014	98.8	78	78	
IES51B3931_mic	gi|76589367	scaffold51_143	35050	99.5	416	416	
IES51A15637_mic	GB	scaffold51_106	297971	98	47	47	
alpha51D_IES2_mic	gi|1752672	scaffold51_159	206601	96.8	29	28	

Table S3. Validation using previously characterized IESs. Known *P. tetraurelia* strain 51 IESs were taken from GenBank. The IESs with no GenBank Accession Number had been communicated to M.B. and are designated "GB" to indicate that they are included in the compilation in [1]. Known IESs were compared with the genome-wide set of IESs by BLASTN. The four IESs that were not found are in bold. Sequencing errors in IESs deposited in GenBank, or SNPs between different 51 stocks, probably explain identity scores below 100%.

Reference

1. Gratias A, Bétermier M (2001) Developmentally programmed excision of internal DNA sequences in *Paramecium aurelia*. Biochimie 83: 1009–1022.

IES (ParameciumDB Accession)	Size (nt)	Location of IES	Gene ohnolog(s)	Ohnologous IES	Element	Match %ID	Match Length (nt)
IESPGM.PTET51.1.98.309432	2003	intergenic	-	none	A	72	883
IESPGM.PTET51.1.128.254421	3392	GSPATG00032293001	none	none	A	76	2981
IESPGM.PTET51.1.77.209216	4154	intergenic	-	none	A	75	3714
IESPGM.PTET51.1.103.177611	2462	GSPATG00028199001	Recent WGD	none	B	87	2464
IESPGM.PTET51.1.104.49056	2238	GSPATG00028299001	none	none	B	89	2228
IESPGM.PTET51.1.105.239361	2483	intergenic	-	none	B	80	2166
IESPGM.PTET51.1.120.182371	1769	GSPATG00031026001	Intermediate WGD	none	B	80	1648
IESPGM.PTET51.1.132.167159	1500	intergenic	-	none	B	86	1465
IESPGM.PTET51.1.163.702	1956	GSPATG00036746001	none	none	B	74	1209
IESPGM.PTET51.1.169.56908	3272	GSPATG00037429001	Recent WGD	none	B	87	3272
IESPGM.PTET51.1.173.70900	2714	intergenic	-	none	B	67	1466
IESPGM.PTET51.1.174.130670	3001	GSPATG00037898001	Recent WGD	none	B	75	2955
IESPGM.PTET51.1.181.1750	3001	GSPATG00038327001	none	none	B	74	2411
IESPGM.PTET51.1.214.11549	2317	GSPATG00038730001	none	none	B	81	2159
IESPGM.PTET51.1.24.100577	1340	GSPATG00009289001	none	none	B	75	1265
IESPGM.PTET51.1.28.457973	2468	intergenic	-	none	B	68	895
IESPGM.PTET51.1.29.290535	1722	GSPATG00010910001	Recent WGD	none	B	78	1695
IESPGM.PTET51.1.35.111752	2820	GSPATG00012627001	none	none	B	85	2692
IESPGM.PTET51.1.42.397702	2473	GSPATG00014783001	Recent WGD	none	B	84	2263
IESPGM.PTET51.1.42.72890	3389	intergenic	-	none	B	89	3389
IESPGM.PTET51.1.47.408041	2125	GSPATG00016182001	Intermediate WGD	none	B	79	2125
IESPGM.PTET51.1.50.348282	1251	intergenic	-	none	B	81	1150
IESPGM.PTET51.1.51.131273	2219	GSPATG00017060001	Recent WGD	none	B	76	1245
IESPGM.PTET51.1.57.48117	1257	GSPATG00018530001	Recent WGD	none	B	82	1257
IESPGM.PTET51.1.76.220822	3048	GSPATG00022960001	Recent WGD	none	B	84	2933
IESPGM.PTET51.1.77.311405	3470	intergenic	-	none	B	85	1254
IESPGM.PTET51.1.80.84925	1513	GSPATG00023746001	none	none	B	82	1515
IESPGM.PTET51.1.85.45587	3479	GSPATG00024746001	none	none	B	78	2786

Table S4. IESs used to identify the *Anchois* transposon family. The BLASTN match of each IES with the final consensus of the A or B element, as indicated, is given in the last 2 columns of the table (cf. Text S1 for the alignments used to reconstitute the final *AnchoisA* and *AnchoisB* consensus sequences). Match length is the sum of non-overlapping HSPs. Match %ID is the weighted average percent identity of the HSPs.

IES category	Peak	Sizes (nt)	Number of IESs	3n	3n+1	3n+2
Non-coding	**-**	**-**	**10304**	**3481 (33.78%)**	**3101 (30.1%)**	**3722 (36.12%)**
	1	0-34	2725	836 (30.68%)	753 (27.63%)	1136 (41.69%)
	2	*34-42*	*93*	*40 (43.01%)*	*22 (23.66%)*	*31 (33.33%)*
	3	42-52	1509	589 (39.03%)	348 (23.06%)	572 (37.91%)
	4	52-63	843	305 (36.18%)	250 (29.66%)	288 (34.16%)
	5	63-73	793	252 (31.78%)	289 (36.44%)	252 (31.78%)
Coding stopwith	**-**	**-**	**11205**	**3700 (33.02%)**	**3515 (31.37%)**	**3990 (35.61%)**
	1	0-34	2231	728 (32.63%)	616 (27.61%)	887 (39.76%)
	2	*34-42*	*78*	*25 (32.05%)*	*15 (19.23%)*	*38 (48.72%)*
	3	42-52	1391	459 (33%)	368 (26.46%)	564 (40.55%)
	4	52-63	975	321 (32.92%)	297 (30.46%)	357 (36.62%)
	5	63-73	880	267 (30.34%)	330 (37.5%)	283 (32.16%)
Coding stopless	**-**	**-**	**23339**	**7095 (30.4%)**	**7049 (30.2%)**	**9195 (39.4%)**
	1	0-34	10943	3123 (28.54%)	3144 (28.73%)	4676 (42.73%)
	2	*34-42*	*240*	*88 (36.67%)*	*61 (25.42%)*	*91 (37.92%)*
	3	42-52	3510	1108 (31.57%)	936 (26.67%)	1466 (41.77%)
	4	52-63	1834	561 (30.59%)	621 (33.86%)	652 (35.55%)
	5	63-73	1520	457 (30.07%)	545 (35.86%)	518 (34.08%)

Table S5. Deficit of 3n IESs in coding sequences evaluated by peak of the 10 bp periodic size distribution. The distribution of IES length modulo 3 is shown for IESs in non-coding (intron + intergenic) and coding (exon) regions of the genome (cf. Table 4). The coding IESs are further separated according to whether or not they contain a stop codon in phase with the upstream exon sequence. The modulo 3 counts are also presented for each of the first 5 peaks of the 10 bp periodic size distribution; the 2nd "forbidden" peak (see Discussion), containing very few IESs, is in italics.

The 10 bp periodicity in the IES size distribution, a strong constraint probably imposed by IES excision geometry (see Discussion), causes a distortion of the modulo 3 length distribution, both in coding and non-coding sequences. The relative proportion of 3n, 3n+1 and 3n+2 IESs differs from 1/3 for each class and varies among the different peaks of the 10 bp periodicity, leading to an overall excess of 3n+2 IESs. However, comparison of the 3n values for each peak shows in all cases a lower percentage of 3n IESs in coding ("stopless" counts) than in non-coding regions of the genome. Furthermore, there is a lower percentage of 3n "stopless" IESs than 3n "stopwith" IESs for all of the peaks except the second one (in italics), which contains very few IESs.

The two constraints on IES size appear to be independent. First, the same 10 bp periodicity is found both for IESs inserted in coding and in non-coding sequences (Fig. S6). Second, the analysis of each individual peak of the 10 bp periodicity shows that in all cases, the frequency of 3n IESs is lower for stopless IESs in coding sequences compared to IESs in non-coding sequences. The observed deficit of 3n stopless IESs in exons is therefore not an indirect consequence of the 10 bp periodicity in the IES size distribution.

TextS3.

PCR validation of IESs from cluster 5, homologous to solo TIRs of the *Thon* transposon

We used 2 different approaches to validate the size and sequence of IESs with homology to *Thon* solo TIRs. The first approach (A in schema, below) involved PCR primers anchored in MAC-destined sequences flanking the IES, and were used to amplify PGM DNA using the Expand Long Template PCR System (Roche Applied Science) with buffer 3 that contains detergents to help denature foldback structures that might be formed by the palindromic TIR of Thon (cf. TextS2), according to the recommendations of the supplier.

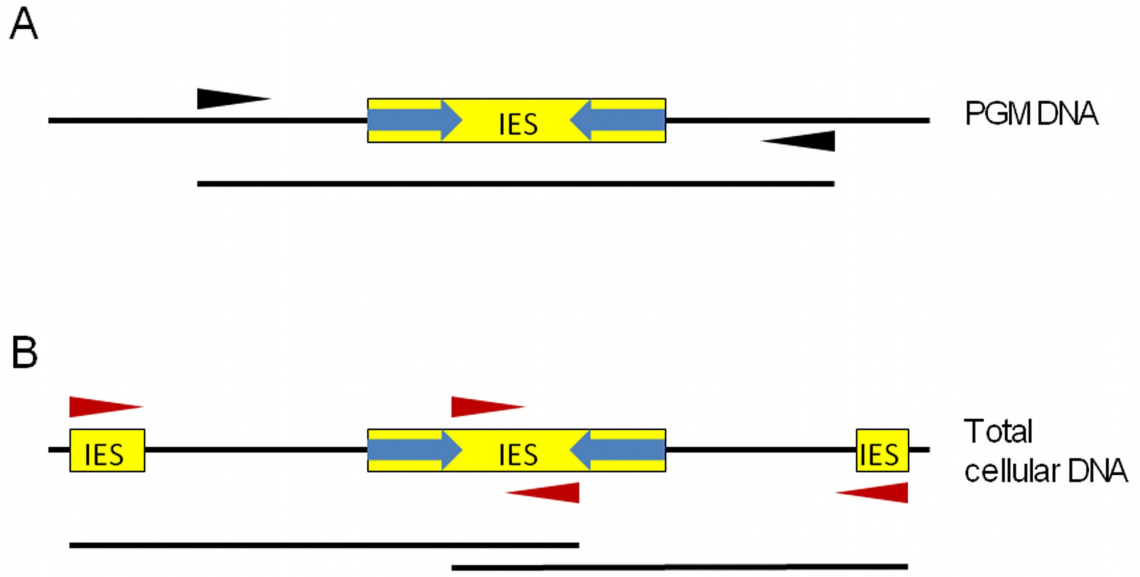
The second approach (B in schema, below) involved PCR primers anchored in flanking IESs, allowing amplification of total cellular DNA, possible for two of the IESs in cluster 5 (cf Table, below). In order to sequence the products, two primers internal to the solo TIR were designed in the central region between the palindromic repeated sequences (cf. Text S2 and schema, below). The same Long Template PCR System was used for the amplifications.

In the case of the three IESs that were completely sequenced, no differences were found with the predicted IES sequences obtained using PGM DNA assembly and the MICA pipeline.

The sequences of the primers are available on request.

Scaffold	Position of IES	Expected Size (nt)	Observed Size (nt)	IES size	Sequence	PCR Method
scaffold51_109	40673	921	~900	571	ND	A
scaffold51_128	266698	1131	~1100	689	Confirmed IES	B
scaffold51_131	262422	980	~950	630	ND	A
scaffold51_18	127217	1189	~1100	770	Confirmed IES	B
scaffold51_34	280841	925	~950	512	Confirmed IES	A
scaffold51_58	302214	1006	~1000	640	ND	A

Table of PCR and sequencing results for cluster 5 IESs. ND, not done. The positions of IES insertion are relative to the MAC strain 51 reference genome.



Schema. PCR methods A and B. The blue arrows represent the palindromic repeats of *Thon* TIRs.

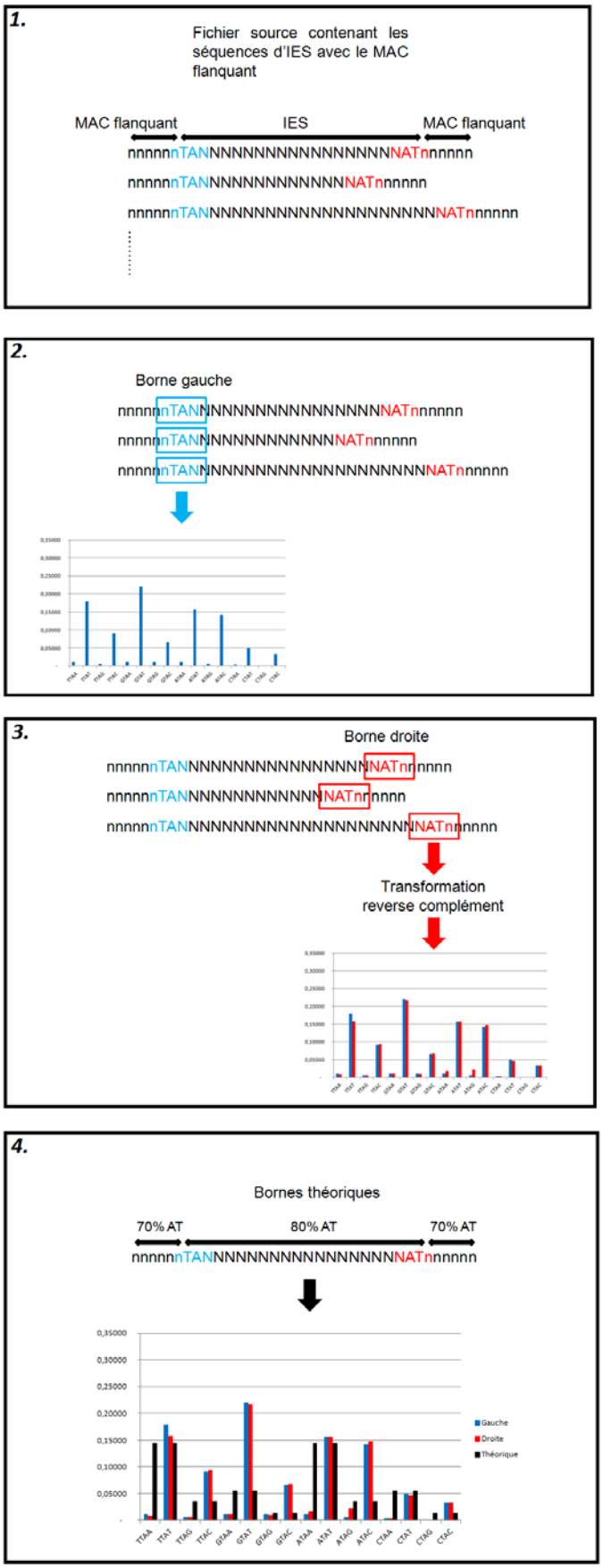


Figure 32 : Procédure pour le classement des séquences aux bornes des IES.

Résultats supplémentaires sur l'analyse des IES

Une fois un jeu de référence des IES obtenu grâce au séquençage de l'ADN de cellules déplétées pour le gène *PGM*, la tentation était grande de regarder plus précisément et à l'échelle globale les biais de séquences qui pouvaient être observés dans les séquences d'IES. C'est ce qui a été fait en partie dans le papier présenté précédemment.

Je voulais pour ma part observer l'abondance aux bornes des IES de la séquence TTAA, qui est la séquence reconnue par la transposase de *piggybac*. Pour cela j'ai écrit un programme pour classer les 45 000 IES selon les bornes observées de chaque coté des IES, en prenant en compte les nucléotides de chaque coté du TA central (figure 32).

J'ai également calculé la proportion théorique d'IES qui devraient avoir une séquence donnée si les nucléotides de chaque coté du TA sont déterminés aléatoirement, en tenant compte toutefois du faible taux de GC dans les séquences MAC de la paramécie et dans les séquences IES. C'est ce qui apparait en noir sur les histogrammes ci contre.

A l'échelle globale

On observe dans un premier temps (figure 33) lorsque l'on considère toutes les IES que certaines séquences sont clairement sous représentées aux bornes des IES, notamment la séquence TTAA qui m'intéressait au départ et qui est la séquence où sont introduites les cassures par la transposase piggybac. Dans l'ensemble toutes les séquences de type XTAA sont sous représentées. A l'inverse des séquences sont sur représentées. Les cas les plus frappants étant les séquences GTAT et ATAC.

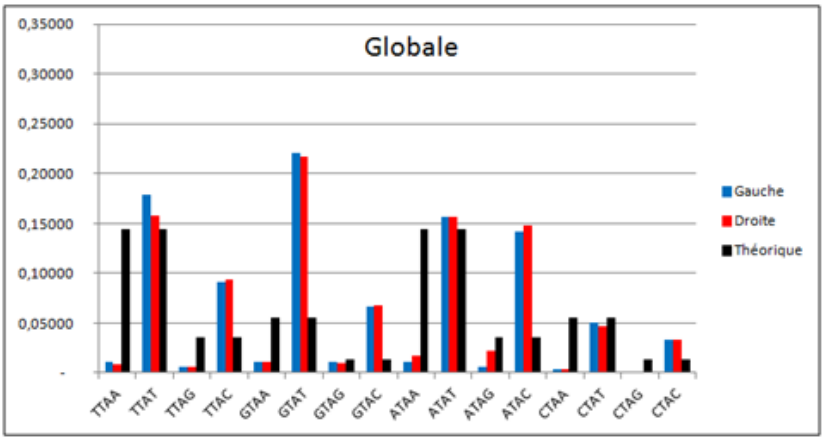


Figure 33 : Analyse globale des séquences d'introduction des cassures double brin programmées aux bornes de toute les IES. En bleu et rouge les fréquences observées pour les bornes gauche et droite, gauche et droite sont déterminées arbitrairement dans le fichier de séquences source.

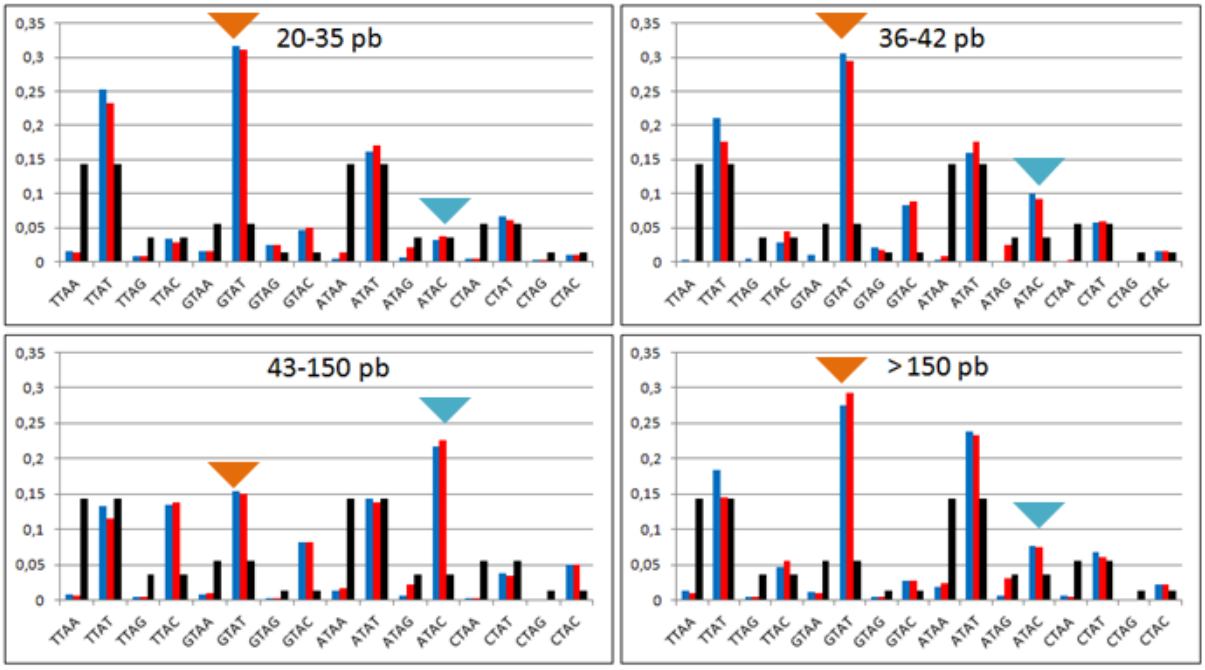


Figure 34 : Analyse des séquences d'introduction des cassures double brin programmées aux bornes des IES en fonction de leur taille. Les séquences GTAT sont indiquées par une flèche orange. Les séquences ATAC sont indiquées par une flèche bleue.

En fonction de la taille des IES

Pour aller un peu plus loin dans cette analyse, j'ai utilisé le même programme après avoir établi différentes classes d'IES (figure 34) :

- La première classe, les IES de 20 à 35 pb correspond aux IES les plus courtes qui représentent un tiers des IES de *Paramecium.*
- La deuxième, pour les IES de 36 à 43 pb, correspond aux IES dont la taille a été contre sélectionnée au cours de l'évolution.
- La troisième classe correspond aux d'une taille comprise entre 44 et 150 pb.
- La quatrième classe correspond aux IES d'une taille supérieure à 150 pb qui représentent moins de 10 de la totalité des IES.

En observant les résultats obtenus on remarque des biais des séquences aux bornes des IES en fonction de leur taille. En particulier, si l'on considère la classe d'IES ayant une taille entre 44 et 150 pb, on note que la borne GTAT, qui est sur représentée aux bornes des IES, est significativement moins sur représentée pour cette classe d'IES. A l'inverse la séquence ATAC est beaucoup plus sur représentée que dans les autres classes d'IES.

Si l'on considère les IES ayant une taille entre 20 et 35pb, on remarque que ces IES présentent plus souvent une séquence TTAT à leur borne.

Les IES d'une taille supérieure à 150 pb ont-elles plus souvent des séquences ATAT présentes à leurs bornes, plus que ce qui est attendu si la séquence est déterminée aléatoirement.

Conclusions

Il ne s’agit que d’une analyse préliminaire mais constitue un exemple de l’apport fourni par le séquençage haut débit des IES. Peut-être découvrira-t-on que ces biais de séquences correspondent à des catégories d’IES qui vont être plus ou moins dépendantes des facteurs d’excision des IES, ou caractéristiques d’un mécanismes d’excision et d’une contrainte topologique particulière.

PAPIER LIGIV-XRCC4 + résultats supplémentaires

Introduction des cassures et maturation

des extrémités cassées d'ADN

Lors de l'analyse du rôle des protéines LigIV et XRCC4 dans les réarrangements programmés du génome de *Paramecium tetraurelia*, j'ai été amené à m'intéresser plus particulièrement à la maturation des extrémités de l'ADN. Il ne s'agit pas du cœur de mon projet qui est l'analyse du rôle de Ku dans les réarrangements programmés du génome mais cela a permis d'apporter une meilleure compréhension des mécanismes moléculaires à l'œuvre pendant les étapes de réparation par la voie NHEJ lors de l'excision des IES.

Introduction des cassures et réparation par le NHEJ

En conditions normales il est possible de détecter avec des techniques de biologie moléculaire fines telle que la « ligation mediated PCR » deux types d'extrémité double brin lors de l'excision des IES. Le premier correspond à la géométrie classique d'une cassure introduite par une transposase de la famille *piggybac*, à savoir 4 bases sortantes en 5'. L'autre type d'extrémité détectée est une cassure avec 3 bases sortantes en 5' qui pourrait correspondre à la forme maturée de l'extrémité d'ADN. En effet les extrémités d'ADN de la séquence MAC flanquante, bien que de géométrie compatible, n'ont pas toujours le dernier nucléotide en 5' complémentaire. Celui-ci doit donc être éliminé pour permettre la maturation de l'extrémité. Alors le modèle postule que les extrémités sont alignées au niveau du dinucléotide TA central. Le gap est alors comblé par une activité de polymérisation avant l'étape finale de ligation.

Maturation des cassures programmées de l'ADN

La maturation des cassures programmées de l'ADN peut être étudiée au niveau de l'extrémité sortante en 5' par cette technique de ligation mediated PCR. En parallèle, pour connaitre le

destin de l'extrémité 3', j'ai réalisé des expériences d'extension à la terminale transférase qui permettent après hybridation d'un adaptateur et amplification par PCR de détecter et éventuellement séquencer ces extrémités 3' d'ADN sans spécificité de séquence et d'adaptateur.

Les résultats obtenus pendant l'étude de la Ligase IV et de son partenaire XRCC4 ainsi que leur implication sur notre connaissance de la maturation des extrémités sont présentées dans cette partie

OPEN ACCESS Freely available online

PLoS GENETICS

Highly Precise and Developmentally Programmed Genome Assembly in *Paramecium* Requires Ligase IV–Dependent End Joining

Aurélie Kapusta[1,2,3☯], Atsushi Matsuda[4☯¤], Antoine Marmignon[1,2,3], Michael Ku[4], Aude Silve[1], Eric Meyer[5], James D. Forney[4], Sophie Malinsky[5,6], Mireille Bétermier[1,2,3]*

1 CNRS UPR3404, Centre de Génétique Moléculaire, Gif-sur-Yvette, France, **2** Université Paris 11, Département de Biologie, Orsay, France, **3** CNRS FRC3115, Centre de Recherches de Gif–sur-Yvette, Gif-sur-Yvette, France, **4** Department of Biochemistry, Purdue University, West Lafayette, Indiana, United States of America, **5** Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, Paris, France, **6** Université Paris Diderot – Paris 7, UFR des Sciences du Vivant, Paris, France

Abstract

During the sexual cycle of the ciliate *Paramecium*, assembly of the somatic genome includes the precise excision of tens of thousands of short, non-coding germline sequences (Internal Eliminated Sequences or IESs), each one flanked by two TA dinucleotides. It has been reported previously that these genome rearrangements are initiated by the introduction of developmentally programmed DNA double-strand breaks (DSBs), which depend on the domesticated transposase PiggyMac. These DSBs all exhibit a characteristic geometry, with 4-base 5′ overhangs centered on the conserved TA, and may readily align and undergo ligation with minimal processing. However, the molecular steps and actors involved in the final and precise assembly of somatic genes have remained unknown. We demonstrate here that Ligase IV and Xrcc4p, core components of the non-homologous end-joining pathway (NHEJ), are required both for the repair of IES excision sites and for the circularization of excised IESs. The transcription of *LIG4* and *XRCC4* is induced early during the sexual cycle and a Lig4p-GFP fusion protein accumulates in the developing somatic nucleus by the time IES excision takes place. RNAi–mediated silencing of either gene results in the persistence of free broken DNA ends, apparently protected against extensive resection. At the nucleotide level, controlled removal of the 5′-terminal nucleotide occurs normally in *LIG4*-silenced cells, while nucleotide addition to the 3′ ends of the breaks is blocked, together with the final joining step, indicative of a coupling between NHEJ polymerase and ligase activities. Taken together, our data indicate that IES excision is a "cut-and-close" mechanism, which involves the introduction of initiating double-strand cleavages at both ends of each IES, followed by DSB repair *via* highly precise end joining. This work broadens our current view on how the cellular NHEJ pathway has cooperated with domesticated transposases for the emergence of new mechanisms involved in genome dynamics.

Citation: Kapusta A, Matsuda A, Marmignon A, Ku M, Silve A, et al. (2011) Highly Precise and Developmentally Programmed Genome Assembly in *Paramecium* Requires Ligase IV–Dependent End Joining. PLoS Genet 7(4): e1002049. doi:10.1371/journal.pgen.1002049

Editor: Gregory P. Copenhaver, The University of North Carolina at Chapel Hill, United States of America

Received October 22, 2010; **Accepted** February 25, 2011; **Published** April 14, 2011

Funding: This work was supported by: Centre National de la Recherche Scientifique (CNRS) (ATIP Plus grant to M Bétermier) (http://www.cnrs.fr/), Agence Nationale de la Recherche (grant BLAN08-3_310945 to M Bétermier and E Meyer) (http://www.agence-nationale-recherche.fr/), National Science Foundation (grant MCB-9506009 to JD Forney)(http://www.nsf.gov/), Ministere de l'Enseignement Superieur et de la Recherche (PhD fellowships to A Kapusta and A Marmignon) (http://www.enseignementsup-recherche.gouv.fr/), Fondation pour la Recherche Medicale (PhD fellowship to A Kapusta and Equipe FRM grant to E Meyer) (http://www.frm.org/), and an EMBO short-term fellowship to A Kapusta (http://www.embo.org/programmes/fellowships/short-term.html). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mireille.betermier@cgm.cnrs-gif.fr

☯ These authors contributed equally to this work.

¤ Current address: Kobe Advanced ICT Research Center, National Institute of Information and Communication Technology, Nishi-ku, Kobe, Japan

Introduction

Although double-strand breaks (DSBs) are among the most deleterious DNA lesions, essential cellular processes involve the programmed introduction of DSBs and their subsequent repair by various pathways. During meiosis, double-strand DNA cleavage mediated by the topoisomerase II-like Spo11 endonuclease triggers homologous recombination, which is essential for the correct segregation of homologs and the mixing of parental alleles without DNA loss (see [1] for review). During lymphocyte differentiation in vertebrates, V(D)J recombination drives the generation of immunoglobulin and T-cell receptor diversity (reviewed in [2]). Programmed DSBs are introduced at both ends of non-coding intervening sequences by the Rag1 domesticated transposase [3] and its partner, Rag2 (reviewed in [4]). Assembly of coding segments then relies on non-homologous end joining (NHEJ) (reviewed in [5]): in this DSB repair pathway, binding of the Ku70/Ku80 heterodimer to DNA ends facilitates their synapsis and recruits other factors involved in their processing and ligation. In the final step, Ligase IV and its partner Xrcc4p are required for covalent joining of the two broken ends.

In ciliates, massive genome rearrangements initiated by developmentally programmed DSBs are associated with DNA elimination during nuclear differentiation [6,7]. In these unicel-

Author Summary

Double-strand breaks (DSBs) are among the most deleterious lesions that may occur on DNA. Some physiological processes, however, involve the introduction of DSBs and their subsequent repair. In the ciliate *Paramecium*, programmed DSBs initiate the extensive genome rearrangements that take place at each sexual cycle, during the development of the somatic nucleus. In particular, short intervening germline sequences (one every 1–2 kb along the genome) are spliced out from coding and non-coding regions. In this study, we present evidence that this process is a two-step mechanism and involves DNA cleavage at both ends of each excised sequence, followed by DSB repair. We demonstrate that cellular end-joining proteins, Ligase IV and its partner, Xrcc4p, are essential for the closure of broken excision sites, which has to be precise at the nucleotide level to allow the assembly of functional genes. This precision stands in sharp contrast to the notion that end joining is an error-prone DSB repair pathway. Therefore, *Paramecium* provides an excellent model for analysis of an intrinsically precise end joining pathway that has been recruited for genome-wide DSB repair.

lular eukaryotes, two kinds of nuclei coexist in the same cytoplasm [8]: the somatic macronucleus (MAC) is essential for gene expression but is destroyed at each sexual cycle, while the germline micronucleus (MIC) undergoes meiosis and transmits its genome to the zygotic nucleus. New MICs and MACs of sexual progeny differentiate from copies of the zygotic nucleus and extensive genome rearrangements take place in the new MAC during this process. In *Paramecium tetraurelia*, MAC development involves DNA amplification (from 2n to 800–1000n) and elimination of two types of germline sequences. Regions extending up to several kb, often including repeated DNA, are eliminated in an imprecise manner, leading to chromosome fragmentation or to internal deletions [9]. In addition, an estimated 60,000 single-copy, short and non coding Internal Eliminated Sequences (IES) are spliced out during assembly of functional genes (reviewed in [10]). *Paramecium* IESs are invariably flanked by two TA dinucleotides, one copy of which is left at their excision site on MAC chromosomes. Their excision is initiated by 4-bp staggered DSBs centered on these conserved TAs [11] and PiggyMac, a domesticated transposase from the *piggyBac* family, is essential for DNA cleavage [12]. Given the predicted density of IESs in the genome, excision would lead to around one DSB every 1–2 kb within the developing MAC.

The question addressed in this study is how *Paramecium* processes DSBs to achieve the precise assembly of somatic genes. Careful examination of cleaved IES ends has led to a mechanistic model, in which two DSBs, one at each end, initiate IES excision [11]. Genetic evidence for a crosstalk between ends before DNA cleavage further supported the view that IES ends are recognized or cleaved in a concerted manner [13]. It was proposed, therefore, that an end-joining DSB repair activity carries out the closure of excision sites on MAC chromosomes. However, alternative models involving a single initiating DSB (at either end) followed by DNA transesterification, have not been definitively ruled out. Indeed, molecules with a single cleaved end are detected in *P. tetraurelia* [13]: nucleophilic attack of the other IES boundary by the free flanking DNA could directly assemble a MAC junction and liberate a linear IES, as proposed for the related ciliate *Tetrahymena* [14]. Alternatively, attack from the free IES end would directly circularize the IES and leave a DSB at the excision site, as suggested for the more distant ciliate *Oxytricha* [15]. In *Paramecium*, linear and circular IES molecules are produced during MAC development [16], making it difficult to draw firm conclusions in favor of any model. Identifying the pathway(s) involved in the last steps of IES excision (assembly of chromosome and circle junctions) should provide new insight into how initial cleavages are introduced at IES ends.

We report here that an end-joining pathway is required for IES excision. Focusing on ATP-dependent DNA ligases, we identified nine genes in *P. tetraurelia*, two of which encode homologs of Ligase IV, an essential actor of the NHEJ pathway. We also found one *XRCC4* homolog. *LIG4* and *XRCC4* genes are specifically induced during sexual processes, before MAC development starts, and a Lig4p-GFP fusion localizes to the developing new MAC, in which IES excision takes place. Functional inactivation of *LIG4* or *XRCC4* completely abolishes the formation of chromosome and circle junctions, demonstrating that IES excision is initiated by two DSBs introduced at both ends of each IES, followed by Ligase IV/Xrcc4p-dependent repair. We propose that the remarkable precision in end joining is largely driven by the characteristic structure of the broken ends generated during IES excision.

Results

ATP-dependent DNA ligases in *P. tetraurelia*

As a first step towards understanding the role of end joining pathways in IES excision, we searched for ATP-dependent DNA ligases encoded by the *P. tetraurelia* MAC genome and found nine putative genes, based on sequence homology. Some of them grouped in pairs with high nucleotide sequence identity (Figure 1), as a consequence of the most recent whole genome duplication (WGD) that occurred during evolution of the *P. aurelia* group of species [17]: these duplicated genes will be designated as ohnologs.

Three major families of ATP-dependent DNA ligases have been described in eukaryotes [18]. Type I ligases (Lig1p) are mainly involved in the ligation of discontinuous Okasaki fragments during replication and in the repair of DNA nicks. Type III ligases (Lig3p) are specialized in the repair of single-strand lesions in the nucleus and in mitochondria: they are restricted to metazoa and Lig1p can perform their function in other organisms. Finally, type IV ligases (Lig4p) are strictly essential for DSB repair *via* the NHEJ pathway. To classify *Paramecium* ligases, we constructed a phylogenetic tree with 52 other ATP-dependent DNA ligases from various organisms, including prokaryotes (Figure S1). We readily identified three *Paramecium* Lig1p and two Lig4p homologs but no clear Lig3p (Figure 1). The last four *P. tetraurelia* ligases formed a monophyletic group with type K ligases, diverged ligases first identified in kinetoplastid protozoan parasites, where they are involved in maintenance of mitochondrial kDNA [19,20].

Alternative end-joining pathways have been reported and differ with regard to the precision of junction formation (reviewed in [21]). Microhomology-mediated end joining (MMEJ) involves end resection and may use Lig1p or type K ligases [22] to generate heterogeneous junctions. In contrast, the canonical NHEJ pathway is able to repair DSBs precisely, without any nucleotide loss, and relies on Lig4p. Because *Paramecium* IES excision is highly precise, we concentrated on this pathway and searched the *P. tetraurelia* genome for genes encoding Xrcc4p, the other essential component of the NHEJ ligation complex [23]. Thorough *in silico* analyses identified one *XRCC4* homolog, in spite of low primary sequence conservation (Figure S2). We also found two putative *XLF/Cernunnos* genes (Figure S3), which encode a cofactor of the Lig4p-Xrcc4p complex [24,25]. We focused our work on *LIG4*

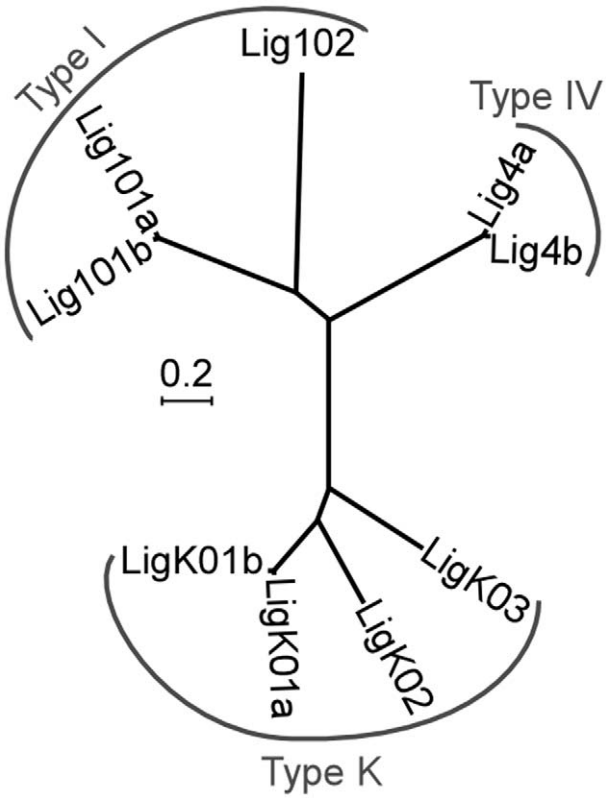


Figure 1. Neighbor-joining tree of *P. tetraurelia* ATP-dependent DNA ligases. Ligases produced from ohnologous genes are designated as ''a'' and ''b''. The tree is based on protein sequences using the following parameters: bootstrap 1000, pairwise deletion of gaps, Poisson correction, uniform rates among sites. Accession numbers in ParameciumDB: *LIG101a* (*GSPATG00024948001*), *LIG101b* (*PTETG9500001001*), *LIG102* (*GSPATG00030449001*), *LIG4a* (*PTETG5400008001*), *LIG4b* (*PTETG7200002001*), *LIGK01a* (*GSPATG00034046001*), *LIGK01b* (*GSPATG00037262001*), *LIGK02* (*PTETG100004001*), *LIGK03* (*GSPATG00025612001*). The scale indicates the number of amino acid substitutions at each site between related proteins.
doi:10.1371/journal.pgen.1002049.g001

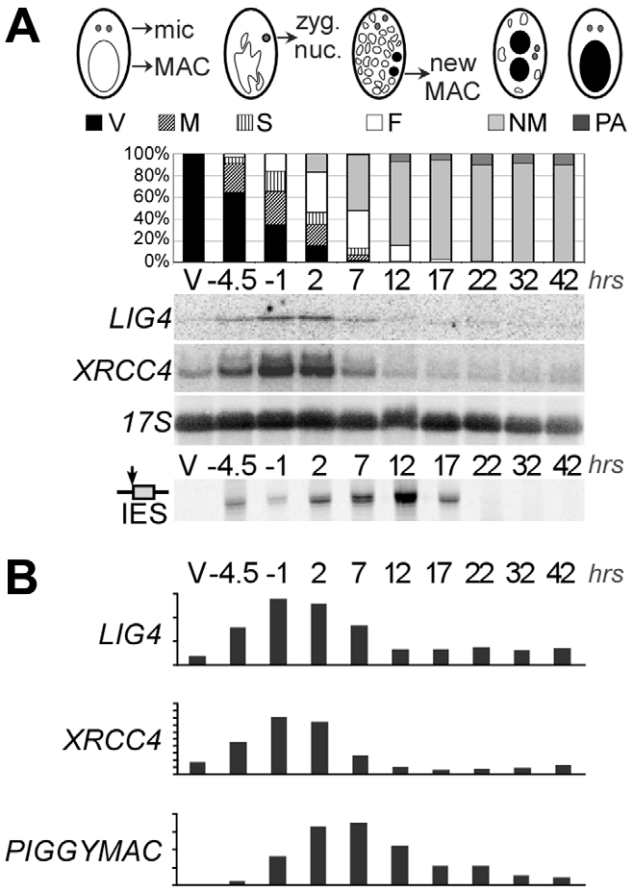


Figure 2. Induction of *LIG4* and *XRCC4* transcription during autogamy. A. Northern blot hybridization of total RNA extracted from strain 51 during an autogamy time-course in standard growth medium. The blot was hybridized successively with *LIG4*, *XRCC4* (see maps in Figure 4A) and 17S rDNA probes. The histograms show the progression of autogamy, with the different cellular stages diagrammed on top. Bottom panel shows the LMPCR-mediated detection of DSBs at the left boundary of IES sm19-576 (broken MAC ends). V: vegetative cells. The following time-points are indicated in hrs. Time 0 corresponds to 50% of the cells harboring a fragmented old MAC. B. Quantification of mRNA levels from the blots shown in A. For *PiggyMac,* values were calculated from a previous hybridization of the same blot [12]. 17S rRNA signal was used for normalization. Y-axes are in arbitrary units.
doi:10.1371/journal.pgen.1002049.g002

and *XRCC4*, because null mutations in these two genes result in embryonic lethality in mouse [26-28].

Developmentally programmed expression of *LIG4* and *XRCC4*

To monitor the transcription of *LIG4* genes, we hybridized northern blots with a *LIG4a* probe. This fragment shared 91% identity with the corresponding region in *LIG4b* and therefore revealed both *LIG4* ohnologs. We detected basal levels of mRNA of the expected size in vegetative cells (Figure 2A). During autogamy, a self-fertilization sexual process, transient accumulation of *LIG4* transcripts was clearly observed at early time-points. The same induction pattern was observed for *XRCC4* and accumulation of both transcripts preceded that of *PiggyMac* mRNA, which encodes the putative IES excisase (Figure 2B). Because northern blot hybridization would not distinguish between *LIG4a* and *LIG4b* transcripts, we performed RT-PCR experiments to distinguish the expression of both genes: only *LIG4a* mRNA was consistently detected by RT-PCR in all experiments and found to be induced early during autogamy, while *LIG4b* transcripts were detected at lower levels and exhibited more variable patterns of expression (not shown). This suggests that most of the signal detected on northern blots may be attributed to *LIG4a* transcription. Early induction of *LIG4a* was clearly confirmed by the statistical analysis of microarrays from four independent autogamy time-course experiments [29], while no significant variation was observed for *LIG4b* throughout early autogamy stages (Figure S3). However, oligonucleotide microarrays do not allow comparing the absolute expression levels of both ohnologs. Microarray analysis also revealed an early induction pattern for *XRCC4* and for one *Cernunnos* ohnolog (*CERa*; see Figure S3).

Sexual reproduction in *Paramecium* includes MIC meiosis, followed by the development of new MACs. To identify more precisely the stage, at which *LIG4* and *XRCC4* are induced, we performed northern blot hybridizations of RNA samples extracted from conjugating cells. Indeed, during conjugation, MIC meiosis takes place within mating pairs and can be clearly distinguished from MAC development, which starts after pair separation [30].

Moreover, conjugation can be synchronized within a 1.5-hr time-window (against 5–6 hrs at best for autogamy), making it possible to separate the two events. For both *LIG4* and *XRCC4* genes, a peak of mRNA was clearly observed 4–6 hrs following mixing (Figure 3A), which corresponds to MIC meiosis, well before exconjugants separate (around 6–7 hrs) and new MACs start to differentiate. Therefore, induction of *LIG4* and *XRCC4* is not a response to DNA double-strand breaks introduced in the new MAC but is rather part of the developmental program of sexual processes.

Lig4p accumulates in developing new MACs during IES excision

We used a *LIG4a-GFP* transgene expressed from the endogenous *LIG4a* promoter to follow the production and localization of the fluorescent fusion protein during synchronized conjugation. *P. tetraurelia* mt7 cells transformed with this transgene were mated with non-injected mt8 cells homozygous for the *pwA* mutation, which produces an abnormal swimming phenotype [31]: after pair separation, exconjugants issued from injected and non-injected cells were sorted according to their swimming behavior. GFP fluorescence was detected in the cytoplasm and MAC of starved injected cells (Figure 3B, panel a). Fluorescence increased during mating (panels b and c) and a strong signal was observed throughout the cells until 13 hrs and concentrated into developing new MACs at 10–13h rs (panels d and e), precisely at the time and place where IESs are massively excised [16]. It returned to background levels when cells resumed vegetative growth (panel f).

During mating, we observed that GFP fluorescence diffused from the injected cell to its non-injected partner (panel c). Therefore, to quantify the overall signal produced from the fusion transgene injected in the mt7 parent, we measured the total GFP fluorescence in vegetative injected cells, in mating pairs, and, separately, in exconjugants produced from injected (wild-type swimming behavior) and non-injected (mutant phenotype) cells. After the separation of exconjugants, the signal measured in non-injected cells (which rapidly lost their fluorescence following pair separation) was added to that of injected cells to calculate the total amount of protein produced by the transgene that was introduced initially in the mt7 parent (Figure 3C). This pointed to a peak of protein accumulation between 10 and 13 hrs following mating, concomitant with the concentration of the GFP fusion into the new MACs.

RNAi against *LIG4* triggers regeneration of the old MAC during conjugation

To investigate the role of Lig4p during conjugation, *LIG4* expression was knocked down by RNA interference (RNAi), which can be obtained by feeding *Paramecium* cells on bacteria induced for the production of double-stranded RNA (dsRNA) [32]. We used plasmid pLIG4b-L to trigger RNAi against both *LIG4a* and *LIG4b*, which are 93.7% identical within the insert carried by this plasmid (Figure 4A). Reactive cells were obtained by starvation in *LIG4*-silencing medium to ensure mRNA level would be knocked down during meiosis, when it was shown to be the highest. Mating pairs were transferred to standard growth medium and the survival of F1 progeny was scored. Although little lethality was observed in the progeny of *LIG4*-silenced cells relative to a control mating (Figure 4B), genetic analysis indicated that survivors were not produced by the successful development of a zygotic MAC. Indeed, mt7 and mt8 parental cell lines were marked with homozygous recessive alleles of *pwB* and *pwA*, respectively, which both give rise to the same mutant swimming phenotype. Authentic

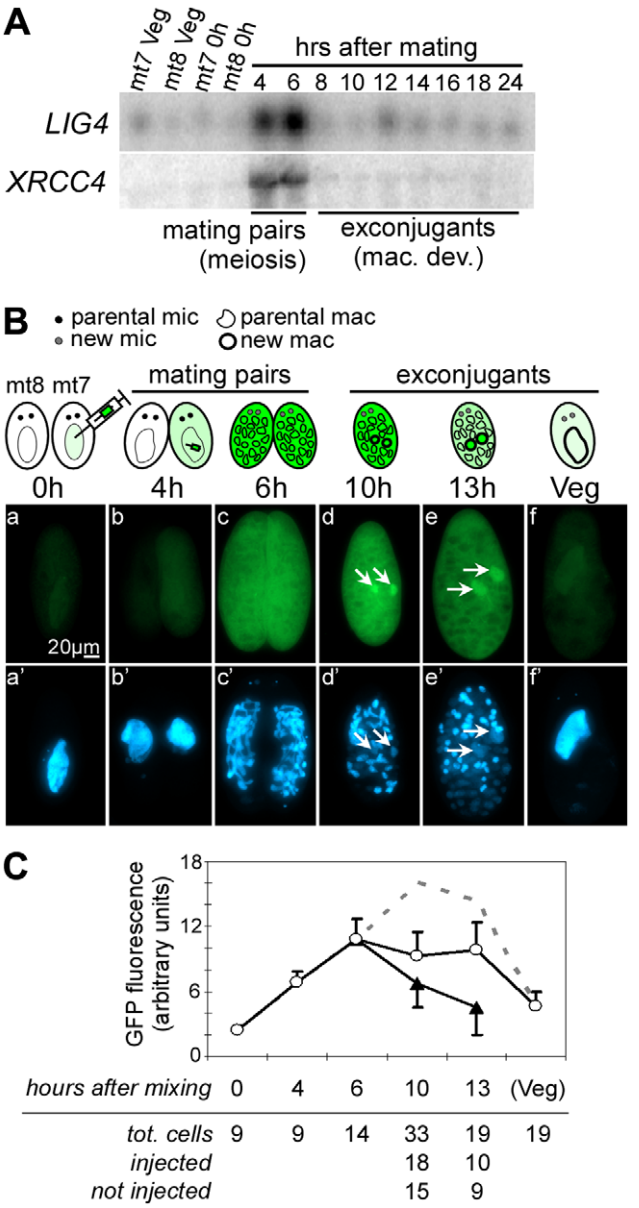


Figure 3. *LIG4* and *XRCC4* expression during synchronized conjugation. A. Northern blot hybridization of total RNA extracted during conjugation of d4-110 mt7 and mt8. *LIG4* and *XRCC4* probes (Figure 4A) were used successively. Mac. Dev.: MAC development. B. Localization of a Lig4a-GFP fusion during conjugation. Reactive mt7 cells transformed with plasmid pLIG401GC, and exhibiting a wild-type swimming behavior (see Materials and Methods), were mated with non injected d4-502 (mt8 *pwA*) at t = 0 hrs. At indicated time-points, cells were fixed and stained with DAPI before observation with a standard fluorescence microscope (GFP filter: a–f; DAPI filter: a'–f'). Developing new MACs are indicated with arrows. C. Quantification of intracellular GFP fluorescence. White circles represent the signal from strongly fluorescent cells originating from the transformed mt7 parent, black triangles represent exconjugants produced by the non-transformed mt8 partner, into which fluorescence has diffused during mating. The calculated total amount of Lig4a-GFP produced by a single transformant is represented as the sum of the two signals (dotted line). Bar = standard deviation.
doi:10.1371/journal.pgen.1002049.g003

F1 progeny were expected, therefore, to be heterozygous at both loci and to have a dominant wild-type phenotype, as observed in the control (Figure 4B). In contrast, all survivors from *LIG4*

silencing exhibited the same mutant phenotype as their parents, as expected if a defect in zygotic MAC development had triggered the regeneration of mutant old MAC fragments [33,34]. Alternatively, a mutant phenotype could also result from conjugation failure with no exchange of gametic nuclei, giving rise to cells with homozygous mutant MICs and MACs. F1 survivors were therefore submitted to an additional round of autogamy in standard growth medium: most survivors (11/15) produced from *LIG4*-silenced parents, as well as those (14/18) obtained from the control (Zyg mac or Regen. in Figure 4B), gave rise to a fraction of wild-type F2 cells and were therefore identified as real F1 exconjugants. A minority of F1 survivors (4/18 in the control, 4/15 in the *LIG4* knock-down), issued from abortive nuclear exchange (NC in Figure 4B), only yielded mutant progeny. Taken together, these experiments show that *LIG4* silencing during conjugation induces parental MAC regeneration in the progeny, indicative of a strong defect in the development of a functional zygotic MAC. Moreover, because new MICs and MACs originate from the MIC of the previous sexual generation, the ability to produce viable F2 sexual progeny from F1 cells with a regenerated old MAC strongly suggests that *LIG4* silencing during meiosis of conjugating cells has not produced deleterious chromosomal alterations in the zygotic nucleus.

LIG4 and *XRCC4* are essential for new MAC development

During conjugation, MAC regeneration is probably favored because there is no active degradation of old MAC fragments when mating pairs are transferred to rich medium [35]. In contrast, during autogamy, cells are kept under prolonged starvation throughout the experiment: this could prevent MAC regeneration due to rapid loss of parental MAC fragments. In *LIG4*-silenced cells, a defect in MAC development might then be monitored by a simple survival assay in the progeny. To test this hypothesis, we injected a constitutive *GFP* transgene into the vegetative MAC of *P. tetraurelia* cells. Injected cells were fed on bacteria producing dsRNA from either of two non overlapping regions of *LIG4* (Figure 4A), then starved to induce autogamy. After 100% autogamy was reached, cells were transferred to standard rich medium to resume vegetative growth (samples I in Figure 4C). A second sample of cells were transferred to rich medium after two additional days of starvation (samples II). Successful formation of a zygotic MAC results in loss of the *GFP* marker contained in the parental MAC, as shown in control experiments. Following *LIG4* silencing, high survival rates were observed when autogamous cells were transferred early to rich medium, but all survivors were fluorescent, indicating that MAC regeneration had occurred. Lethality strongly increased when survival tests were performed after 2 additional days of starvation, consistent with MAC regeneration being prevented by autolysis of old MAC fragments [35].

In all subsequent assays, autogamous cells were kept under prolonged starvation to minimize MAC regeneration, so that we could follow the effect of RNAi on the simple basis of progeny death. Different RNAi plasmids were used to knock down *LIG4a* or *LIG4b* expression: this repeatedly led to massive death of post-autogamous cells (Figure 4C and 4D). In each experiment, MAC regeneration was assayed following starvation of the few survivors in standard medium: *bona fide* sexual progeny were too young to undergo autogamy again, while cells with a regenerated old MAC were able to start a new sexual cycle. The efficiency of each silencing was assayed by quantifying mRNA levels on northern blots. RNAi triggered either by *LIG4a* or *LIG4b* dsRNA resulted in variable decreases in the total amount of *LIG4* mRNA, while a mixture of dsRNA from both genes reduced mRNA levels more

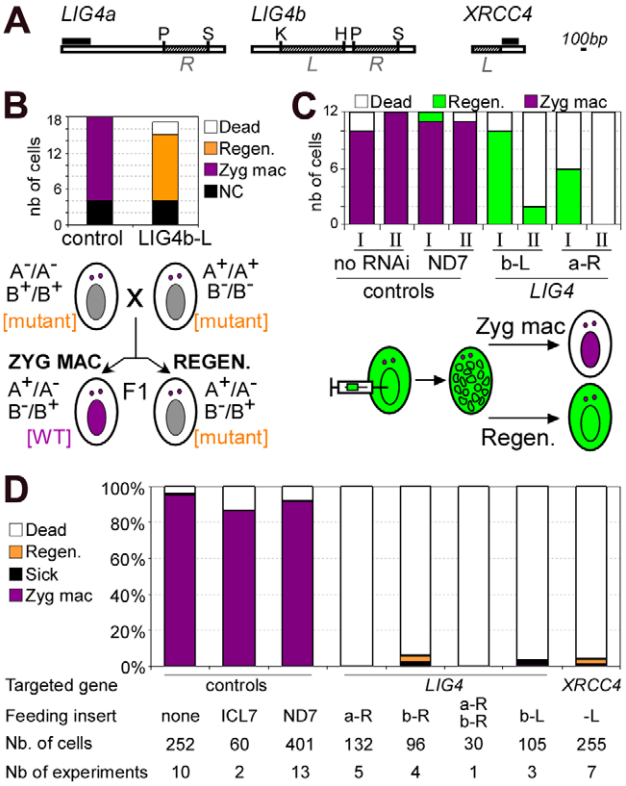


Figure 4. Analysis of sexual progeny of *LIG4*-silenced cells. A. Maps of *LIG4* and *XRCC4* genes. *XRCC4* accession number in ParameciumDB is *GSPATG00029540001*. The fragments used in the construction of RNAi plasmids are shown in hatched boxes and designated in italics (*L* or *R*). Probes used for Northern blot hybridization are displayed as black boxes: we used the first 510 bp of *LIG4a* open reading frame and a 302-bp fragment encompassing bp 508–809 for *XRCC4*. Restriction sites: P = *Pst*I, S = *Spe*I, K = *Kpn*I, H = *Hind*III. B. Survival and genetic analysis of F1 progeny from a cross between a3093 (mt7 *pwB*) and d4-502 (mt8 *pwA*). The histogram shows the number of cells surviving conjugation and the genotype of their MAC, out of 18 cells for the control experiment and 17 for the *LIG4* knock-down. NC: cells resulting from a conjugation failure within a mating pair (mutant phenotype in F1 and F2). Zyg. MAC: cells containing a functional zygotic nucleus. Regen: MAC regenerants in F1 (mutant phenotype in F1 and recovery of wild-type clones in post-autogamous F2). Control RNAi was carried out with an empty vector (no insert). The segregation of *pwA* (A-) and *pwB* (B-) markers in F1 MICs is shown below the histogram. The cell phenotype determined by the MAC is indicated between square brackets. C. MAC regeneration in post-autogamous progeny of 51 cells injected with a GFP transgene. Cells were transferred at day 0 in control (no RNAi or RNAi against *ND7*) or *LIG4*-silencing medium and, following starvation, 100% autogamy was observed at day 2. Survival tests were performed on 12-cell samples at day 2 (I) or day 4 (II). Regen: MAC regenerants, as judged by GFP fluorescence. D. Survival of post-autogamous progeny following *LIG4* or *XRCC4* silencing (experiments are independent of the one presented in B). Regen: MAC regenerants. The inserts cloned in plasmids used for feeding experiments are displayed in A. The histogram represents a compilation of several independent experiments, in each of which the progeny of 30 individual autogamous cells was assayed for survival.

doi:10.1371/journal.pgen.1002049.g004

significantly (Figure S4). Despite these differences, strong lethality was observed under all conditions in the progeny of silenced cells (Figure 4D). We were concerned that the observed phenotypes could be due to non-targeted silencing of some other transcript by siRNAs produced from *LIG4* RNAi plasmids and, therefore, knocked down the expression of the single-copy *XRCC4* gene,

which encodes an essential partner of Lig4p. RNAi against *XRCC4* was found to efficiently reduce its mRNA down to background (Figure S4D) and, like in the previous experiments, strong lethality was observed in the post-autogamous progeny of *XRCC4*-silenced cells (Figure 4D). At the cellular and molecular levels (see below), identical phenotypes were observed upon targeting either *LIG4* or *XRCC4* transcripts.

RNAi against *LIG4* or *XRCC4* leads either to MAC regeneration or to death of sexual progeny. This suggests that the Lig4p/Xrcc4p complex is essential for the development of a functional new MAC. To monitor MAC development, cells were fixed at different time-points during autogamy, then stained with DAPI. Developing new MACs were observed in *LIG4*- or *XRCC4*-silenced cells but their fluorescence remained very faint, even at late time-points, in sharp contrast to the bright signal observed at similar stages in a control RNAi experiment (Figure 5A). To quantify this difference, we used synchronized conjugation to monitor the total DNA content during MAC development in exconjugants stained with propidium iodide. Exconjugants from *LIG4*-silenced cells often contained smaller new macronuclei relative to the control and quantification revealed consistently lower (~40%) DNA amounts in their new MACs, indicative either of a replication defect or of DNA degradation (Figure 5B). The characteristic faint staining of the new MACs in Lig4p- or Xrcc4p-depleted cells was dependent on the presence of wild-type levels of the PiggyMac transposase, the putative endonuclease involved in double-strand cleavage of IES boundaries [12], as shown by the normal aspect of developing new MACs in a *PGM+XRCC4* double knock-down (Figure 5C). These observations strongly suggest that the Lig4p-Xrcc4p repair complex acts on the PiggyMac-dependent DSBs introduced in the new MAC. DNA under-amplification would, therefore, reflect a DSB repair defect in during MAC development.

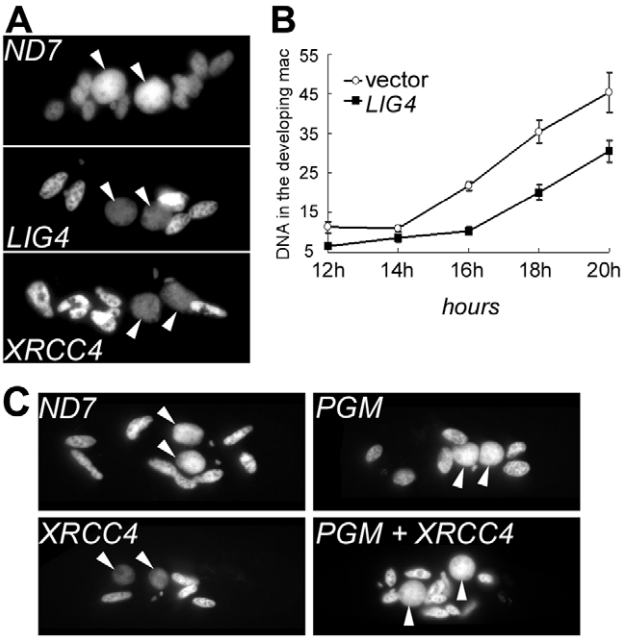


Figure 5. DNA content in developing MACs of *LIG4*- or *XRCC4*-silenced cells. A. DAPI staining of single cells fixed during autogamy in the macronuclear variant 51Δ*A*. Cells were not treated with RNase prior to staining. For *ND7* & *LIG4* silencing, cells were fixed 3 days following transfer to RNAi medium. For *XRCC4,* samples were treated at day 4. White arrows: new MAC. Other stained nuclei are old MAC fragments. B. Quantification of DNA content in propidium iodide-stained exconjugants following RNAi against *LIG4* (pLIG4b-L). Fixed cells were treated with RNase A prior to staining. For each time-point, the average DNA amount (in arbitrary units) contained in the new MACs is displayed in the curve. C. DNA under-amplification in the new developing MACs of *XRCC4*-silenced cells depends on the presence of wild-type levels of PiggyMac transposase. *P. tetraurelia* strain 51 was silenced for the expression of *ND7* (Control), *XRCC4* or *PiggyMac* (*PGM*) individual genes, or cosilenced for *XRCC4* and *PGM*. During autogamy, cells were stained with DAPI and observed using a Zeiss fluorescence microscope (magnification: 630X). Except for the *ND7* control, no viable sexual progeny was recovered from RNAi-treated cells.
doi:10.1371/journal.pgen.1002049.g005

Free broken ends accumulate at IES boundaries in *LIG4* or *XRCC4* knock-downs

To analyze the role of Lig4p-Xrcc4p in DNA rearrangements, we examined IES excision intermediates produced during an autogamy time-course of cells submitted to RNAi against *LIG4* or *XRCC4*.

Total genomic DNA was extracted from vegetative cells and at different time-points during MAC development. We first used LMPCR to detect DSBs at IES boundaries during autogamy of strain 51. Free broken chromosome ends were visualized at early time-points during MAC development of cells submitted to a control RNAi, but they were only transient, indicative of efficient closure of IES excision sites (Figure 6A). In cells silenced for *LIG4*, DSBs started to accumulate at the same autogamy stage but broken ends were found to persist until the last time-points, both on the flanking sequences that should be joined to form MAC DNA, which will be designated as MAC ends (Figure 6A), and at IES ends (Figure 6B). These observations support the hypothesis that *LIG4* silencing results in a defect in DSB repair, downstream of DNA cleavage. Therefore, we followed the closure of IES excision sites during autogamy, using strain 51Δ*A*, a macronuclear variant of strain 51, which harbors a wild-type MIC genome but carries a deletion of the surface antigen *A* gene in its MAC [13]. We took advantage of the fact that, during autogamy of 51Δ*A*, the *A* gene (absent from the parental MAC) is transiently amplified in the new developing MAC, before being deleted at later stages, making it possible to monitor the formation of *de novo* excision junctions by a simple PCR assay. Such junctions were readily amplified from control cells but none could be detected in *LIG4*- (Figure 7A) or *XRCC4*-silenced cells (Figure S7A), indicating that the precise closure of IES excision sites on MAC chromosomes requires Lig4p-Xrcc4p. In these experiments, we confirmed by LMPCR that free broken ends accumulate at IES boundaries in 51Δ*A* cells silenced for *LIG4* (Figure S6) or *XRCC4* (Figure S7B). Interestingly, specifically for IESs from the *A* gene, DSBs with the expected 5′ overhangs were introduced normally at early time-points but could not be detected at later time-points (Figure S6), perhaps as a consequence of maternal inheritance of the *A* gene deletion. With regard to excised molecules, we reported previously that IESs of sufficient length are circularized after excision [16]. We used divergent primers internal to some IESs to monitor the formation of covalently closed excised circles by PCR (Figure 7B). Here again, no junction products could be detected in cells silenced for *LIG4*, indicating that end joining is also required for IES circularization.

Lig4p-Xrcc4p is dispensable for DSB 5′ processing but is required for nucleotide addition to 3′ ends

DSB processing during autogamy is thought to occur before the final joining of chromosome ends at IES excision sites [11]. Processing includes the removal of the 5′ terminal nucleotide and the addition of one nucleotide to the 3′ end of the break.

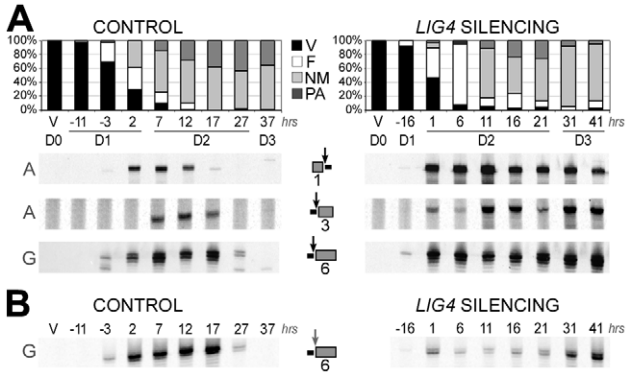


Figure 6. LMPCR detection of free broken ends at IES boundaries in *LIG4*-silenced cells. Three IESs are presented: 1 = 51A1835 (28bp), 3 = 51A4404 (77 bp) from surface antigen *A* gene and 6 = 51G4404 (222 bp) from surface antigen *G* gene. Gene names are indicated on the left of panels A and B. IESs are drawn as grey boxes and MAC flanking sequences as black lines. Vertical arrows indicate the position of DNA cleavage in each experiment. All details about the linkers and primers used in this experiment are provided in Table S1. In both time-courses, cells were transferred to RNAi medium at day 0 (D0). The T0-time point is the time when 50% of cells have a fragmented old MAC. RNAi against *LIG4* was obtained using mixed bacterial cultures producing dsRNA from *LIG4a* (pLIG4a-R) and *LIG4b* (pLIG4b-R). Histograms show the progression of autogamy. V: vegetative cells. F: fragmented parental MAC. NM: cells with two visible new developing MACs. PA: post-autogamous cells with one MAC and surrounding fragments. A. LMPCR detection of free broken MAC ends at IES excision sites during autogamy of strain 51 submitted to control or *LIG4* RNAi. B. LMPCR detection of free broken IES ends at the left boundary of IES #6 during the same time-course.
doi:10.1371/journal.pgen.1002049.g006

To analyze 5′ end processing, LMPCR products obtained following *in vitro* ligation of a linker to the free 5′ overhangs carried by broken MAC ends, can be resolved at the nucleotide scale on denaturing polyacrylamide gels (Figure 8A). This makes it possible to distinguish a doublet of bands, clearly visible at early-time points (6 hrs) in a control autogamy time-course and during autogamy of *LIG4*-silenced cells. Only the bottom band was found to accumulate until the last time-point (40 hrs) in cells depleted for Lig4p. The linker used in this experiment carries a nonphosphorylated 5′-ATAC overhang that guides linker ligation to the 4-base 5′ extension generated at the broken MAC end by initial PiggyMac-dependent cleavage. As previously reported [11], *in vitro* ligation may also occur through a 1-nt gap, which would covalently join the linker to a processed 5′ end, from which the 5′ terminal nucleotide has been removed (see diagrams at the bottom of Figure 8A). The presence of 4-base and 3-base 5′ overhangs was confirmed by DNA sequencing of the doublets observed at early time-points, while only 3-base overhangs accumulate at later time-points in *LIG4*-silenced cells. We can conclude, therefore, that Lig4p is not required for normal removal of the 5′ terminal nucleotide from broken MAC chromosome ends.

To investigate whether the polymerizing step still takes place in *LIG4* knock-downs, we used terminal transferase-mediated poly(C)-tailing to map the position of free 3′ ends at broken MAC ends. For the two IESs presented in this study, we detected only the 3′ end generated by initial PiggyMac-dependent cleavage when cells were depleted for Lig4p, in contrast to the control, in which the poly(C) tail was branched at the expected two positions (Figure 8B and Figure S8). This result indicates that nucleotide addition to broken 3′ ends is strongly impaired in *LIG4*-silenced cells.

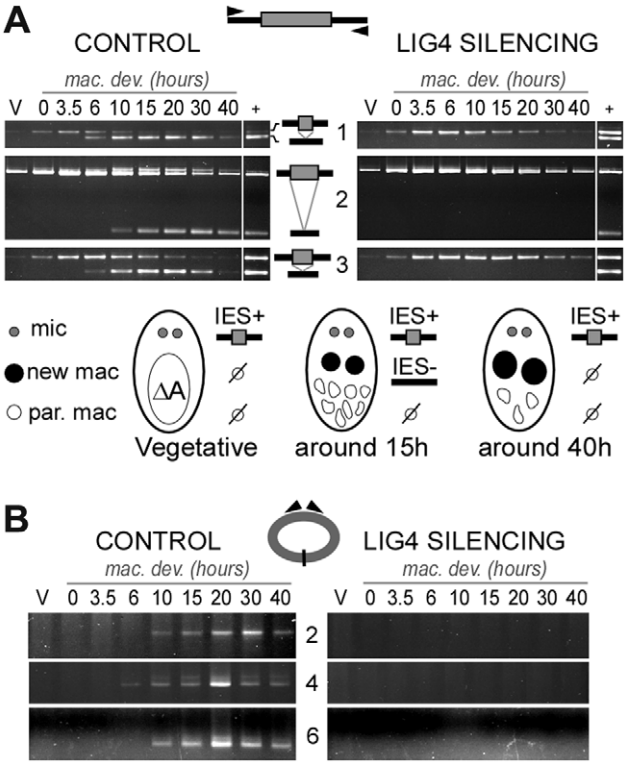


Figure 7. Absence of final junction products following IES excision in *LIG4*-silenced cells. IESs #1 (= 51A1835), 3 (= 51A4404) and 6 (= 51G4404) are described in the legend to Figure 6. Two additional IESs from surface antigen *A* gene are presented: 2 = 51A2591 (370 bp) and 4 = 51A4578 (883 bp). In each diagram, IESs are represented by a grey box and their flanking MAC sequences by a black line. PCR primers are drawn as arrowheads. A. Detection of IES excision junctions in the *A* gene in the macronuclear variant 51Δ*A*. Top: gel electrophoresis of PCR products around IESs. Bottom: representation of MAC development stages. In vegetative cells, only the MIC contributes to the PCR signal (IES+). When new MACs develop, *de novo* chromosomal junctions give an IES- PCR signal. At later stages, this signal disappears due to maternal epigenetic inheritance of the *A* deletion [13]. B. Detection of IES circle junctions by PCR in 51Δ*A*, using two internal divergent primers for each IES (same time-course as in A).
doi:10.1371/journal.pgen.1002049.g007

Discussion

IES excision: a "cut-and-close" mechanism

Ever since developmentally programmed genome rearrangements were reported in ciliates, identifying the key enzymes that catalyze DNA elimination has constituted a challenge. DNA rearrangements were shown to be maternally controlled *via* non-coding RNAs and a specialized RNA interference pathway, which mediate genome-wide comparison of maternal MIC and MAC genomes and guide elimination of MIC-restricted sequences from the new MACs (reviewed in [36,37]). Previous screens for indispensable IES excision genes in *P. tetraurelia* uncovered the essential role of a developmentally regulated SUMO pathway likely to operate in the developing new MAC [34] and of Die5p, a nuclear protein of unknown function acting at a late step during DNA rearrangements [38]. The recent discovery that domesticated transposases are essential for initial DNA cleavage in *Paramecium* and *Tetrahymena* has provided strong evidence that IES excision is related to cut-and-paste transposition [12,39]. In *Oxytricha*, a more distant ciliate, transposases from another family have also been implicated in DNA rearrangements [40].

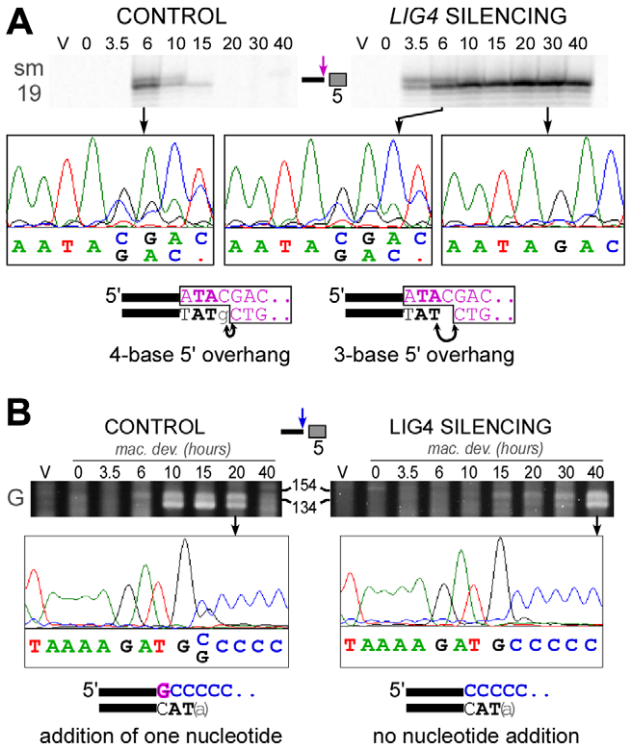


Figure 8. Analysis of 5′- and 3′-end processing at double-strand breaks in *LIG4*-silenced cells. A. LMPCR detection of 4-base and 3-base 5′ overhangs on the broken MAC ends generated at the left boundary of IES #5 = sm19-576 (66 bp) during autogamy of strain 51Δ*A* submitted to control or *LIG4* RNAi. LMPCR products were separated on high resolution polyacrylamide denaturing gels to visualize the doublet of bands. The position of linker ligation to the free 5′ end was confirmed by gel purification of LMPCR products and sequencing using primer sm19-4 specific for the left flanking MAC sequences (the chromatograms show the sequence of the top strand). The structure of the ligation products is diagrammed at the bottom, with the linker represented in purple. Arrows indicate the nucleotides that are ligated to the linker on 4-base or 3-base 5′ extensions. B. Nucleotide addition to broken MAC 3′ ends is impaired in Ligase IV-depleted cells. Terminal transferase-mediated poly(C) tailing of the MAC left 3′ end of IES #6 (51G4404) was performed as indicated in [11]. Poly(C)-tailed products were amplified using primers 51G18 and I, then a nested PCR was performed with 51G13 and I before electrophoresis on a 3% Nusieve agarose gel. The position of size standards (in bp) is indicated: PCR products are expected around 122 bp, with some variability due to the length of the poly(C) tail added by the terminal transferase. To determine the position of poly(C) addition, gel-purified PCR products were sequenced using 51G13 (from the MAC left flanking sequences) as a sequencing primer. The chromatograms show the sequence of the top strand and the structure of the broken MAC ends identified in each sample is displayed at the bottom.
doi:10.1371/journal.pgen.1002049.g008

Our data point to an absolute requirement for Lig4p and Xrcc4p in IES excision, downstream of PiggyMac-dependent cleavage of IES ends. Consistent with a direct participation in DNA rearrangements, a Lig4p-GFP fusion protein accumulates in developing new MACs by the time IES excision takes place. Most IESs studied in this work are cleaved at the same developmental stage in *LIG4*-silenced cells as in the control, suggesting that depleting Lig4p does not interfere directly with DNA cleavage. One notable exception is IES 51G4404, for which a delay in DSB introduction was conspicuous, both by poly(C) tailing (Figure 8B) and by LMPCR analysis of DSBs (Figure S6), in autogamy time-courses with a particularly good synchrony. However, this difference was not observed in all cultures and additional work is needed to evaluate the generality of this observation.

Both the precise closure of excision sites on MAC chromosomes and the circularization of excised IESs require the Lig4p-Xrcc4p end-joining complex. Formally, our data may still fit with a model in which cleavage at one IES boundary would initiate excision, while second-end cleavage would be impaired in cells depleted for Lig4p or Xrcc4p. Single-end cleaved IESs still attached to their flanking MAC sequences at their other end can indeed be detected by LMPCR during the course of autogamy in standard medium [13], or by the tailing of 3′ ends in control or *LIG4* knock-downs (Figure S8B). However, they appear only transiently and do not accumulate in Lig4p-depleted cells, in contrast to DSBs at flanking MAC-destined ends (Figure S8A and S8B). Furthermore, LMPCR experiments using a linker compatible with both ends of one particular IES (51G4404) allowed the detection of larger amounts of excised linear molecules in *LIG4*-silenced cells than in a control RNAi (Figure S8D). These observations rule out the participation of direct DNA transesterification in assembly of chromosome and circle junctions and rather support a model, in which two double-strand cleavages, one at each end, initiate IES excision ([11], see Figure 9A). We propose that Lig4p, in association with Xrcc4p, precisely joins the two resulting broken ends on MAC chromosomes (Figure 9B). Similarly, covalently closed circles are secondary products of the reaction and their formation also depends on Lig4p-Xrcc4p: for those IESs flexible enough to bring their ends together, circularization would prevent reactive 3′OH groups generated by DNA cleavage at IES ends from transposing to other target sites in the genome. Thus, in contrast to cut-and-paste transposition, IES excision uses a "cut-and-close" mechanism, in which developmentally programmed end joining efficiently drives somatic chromosome assembly and avoids reintegration of excised sequences.

In the macronuclear variant 51Δ*A*, which carries a macronuclear deletion of the *A* gene, we observed that, specifically for IESs carried by the *A* gene, DSBs were introduced normally, but they diminished over time (Figure S6), concomitantly with the deletion of the *A* gene. Notably, macronuclear deletion of the *A* gene is associated with chromosome fragmentation, with telomere addition at heterogeneous positions upstream of the gene transcription start [41]. This points to a possible mechanistic difference between IES excision and chromosome fragmentation, and suggests that the breaks that cause imprecise elimination of the *A* gene are not protected, leading to degradation of the whole region.

DSB repair and DNA replication during MAC development

The transcription of *LIG4* and *XRCC4* is induced during meiosis, largely before programmed genome rearrangements take place in developing new MACs. This pattern led us to consider the possibility that NHEJ proteins may be involved in the repair of meiotic DSBs and that a depletion in Lig4p-Xrcc4p could induce deleterious genome rearrangements during MIC meiosis, which would be transmitted to the new MICs of the progeny and strongly impinge on normal development of the new MACs. However, we checked the progression of meiosis by DAPI staining during autogamy of cells silenced for *LIG4* or *XRCC4* expression and observed that meiotic divisions I and II occur normally (Figure S5), with no arrest in meiosis until new MACs differentiate from mitotic copies of the zygotic nucleus. Furthermore, genetic analysis of the progeny of conjugating cells silenced for *LIG4* indicated that the new MICs of the following sexual generation were fully functional germline nuclei (see Figure 4B and related text). We did observe a novel phenotype in the new MACs that develop in cells silenced for *LIG4* or *XRCC4*: quantitative analysis revealed an

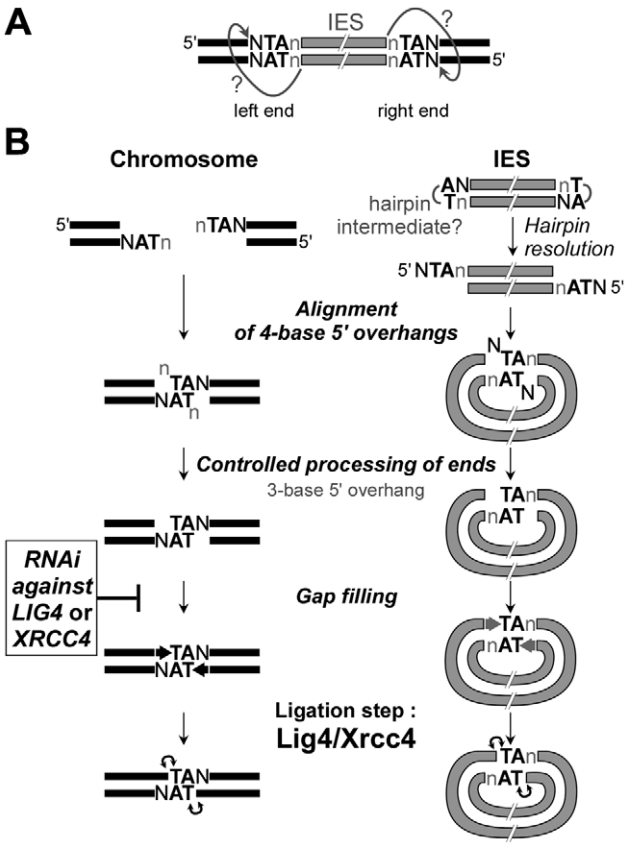


Figure 9. Model for IES excision in *P. tetraurelia*. A. Model for initial cleavage at IES boundaries by the PiggyMac-associated complex. Nicking of the first strand would liberate a reactive 3′OH residue at IES ends. The putative nucleophilic attack of the top strand is represented by arrows with question marks. B. DNA intermediates and end joining factors involved in IES excision are shown. In wild-type conditions, repair of the chromosomal excision sites (left) is probably mediated by alignment of 4-base 5′ overhangs *via* the pairing of their central conserved TA. The formation of IES circles (right) may require the resolution of putative hairpins at IES ends, prior to the formation of intramolecular paired-end intermediates. The actors involved in the controlled processing of ends are still unknown. At least for the MAC chromosome junctions, the removal of the 5′ terminal nucleotide does not require Lig4p/Xrcc4p, in contrast to the gap filling step, which is strongly inhibited in *LIG4* knock-downs. The final ligation is totally Lig4p/Xrcc4p-dependent. In *LIG4* or *XRCC4* silencing, no MAC junctions are formed and DSBs accumulate. No circle junctions could be detected.
doi:10.1371/journal.pgen.1002049.g009

anomalously low DNA content within these nuclei (Figure 5). Strikingly, this phenotype was suppressed by a double RNAi targeting both *PiggyMac* and *XRCC4*, suggesting that it does not result from a meiotic defect but, rather, from a global replication slow-down caused by the accumulation of PiggyMac-dependent DSBs in the developing new MACs, or from active DNA degradation at unrepaired broken ends. Persistent DSBs with the expected geometry could be detected at IES ends using a sensitive LMPCR assay, even at late time-points. This suggests that broken ends are, at least in part, protected against extensive degradation in cells depleted for Lig4p or Xrcc4p and favors the idea that unrepaired DSBs at IES excision sites would block the progression of DNA replication. In spite of all our efforts, we have been unable to detect any degradation products by Southern blot hybridization of total genomic DNA (not shown), consistent with the idea that broken chromosomes are under-amplified.

Free, but incompletely processed broken ends accumulate at IES boundaries in *LIG4* or *XRCC4* knock-downs

Like the PiggyMac domesticated transposase, the Lig4p-Xrcc4p complex is an essential component of the IES excision core machinery. In our "cut-and-close" model for IES excision, the final ligation step performed by the Lig4p-Xrcc4p complex is thought to take place within a paired-end intermediate (Figure 9B), in which the two 4-base 5′ overhangs generated by initial cleavage at each IES boundary are aligned *via* the pairing of their central conserved TA dinucleotides. In this intermediate, the 5′ terminal base of each broken end is generally not complementary to its facing nucleotide and, as reported earlier, highly controlled removal of these mismatched bases by yet unknown nuclease activities, and gap filling by addition of one nucleotide to the free 3′ end precede the final ligation step [11]. Interestingly, close examination of LMPCR products separated on high-resolution denaturing gels revealed that most persisting broken ends in *LIG4*-silenced cells carry a 3-base 5′ overhang (Figure 8A and data not shown). The conversion of 4-base to 3-base overhangs is thought to reflect the removal of the 5′-terminal nucleotide during the repair of IES excision sites ([11], see Figure 9B). Our observation, therefore, suggests that controlled 5′-processing of broken ends still takes place in cells depleted for Lig4p-Xrcc4p. In contrast, our data point to an inhibition of the polymerizing reaction that carries out the addition of one nucleotide to the free 3′ end prior to the ligation step (Figure 8B and Figure S8C). Our *in vivo* data provide further support to the idea that polymerase activity (or recruitment) during NHEJ repair is strongly dependent on Lig4p-Xrcc4p, a hypothesis previously proposed by others, based on observations made in a cell-free NHEJ system [42].

Another implication of our results is that the processed chromosome ends appear to be protected against extensive resection, as judged by the accumulation of LMPCR products observed in cells depleted for Lig4p. However, for a given IES boundary, we noticed an asymmetry in the amounts of broken ends detected on the IES side (ends of excised molecule) and on the MAC side (flanking the excision site). Even though LMPCR assays are not quantitative, free IES ends were consistently detected in lower amounts in *LIG4*-silenced cells relative to control, while their detection level still increased at late time-points (Figure 6B). Two non-mutually exclusive hypotheses may account for this asymmetry. On the first hand, as shown for a canonical *piggyBac* transposase [43], PiggyMac may introduce hairpins at the ends of excised IESs (Figure 9), which would not be detected in our LMPCR assay. These hairpins might be less efficiently converted to 5′ overhangs if recruitment of Lig4p-Xrcc4p were required to stabilize a resolution complex. On the other hand, broken MAC ends, but not IES ends, may be strongly protected against resection, even in the absence of Lig4p-Xrcc4p. Such protection could be mediated by the PiggyMac-associated cleavage complex itself. Alternatively, the Ku70/Ku80 heterodimer, which was shown by others to control the precision of the NHEJ pathway by inhibiting DSB resection [44], could also contribute to protecting chromosome ends. *P. tetraurelia* harbors several *KU* genes in its genome, some of which are specifically induced during MAC development (Figure S3), and future work will elucidate their role in IES excision.

Highly precise NHEJ and a domesticated transposase participate in developmentally programmed genome rearrangements

Our work provides strong evidence that the Lig4p-Xrcc4p complex, a key actor of the NHEJ pathway, carries out DSB repair

during developmentally programmed IES excision in *Paramecium*. End joining is highly precise at the nucleotide level and this is critical for the assembly of functional open reading frames in the somatic genome. Several observations support the idea that tight coupling between DNA cutting and repair contributes to this precision. The early induction of *LIG4* and *XRCC4* genes, well before new MACs start to develop, indicates that their expression is developmentally programmed rather than triggered by DNA breaks: thus, the preexisting end-joining ligation complex could readily be recruited to DSBs as soon as they appear in the new developing MACs. Moreover, we have proposed that IES ends form a synapse before they are cleaved [13], perhaps as a result of binding by the PiggyMac domesticated transposase. If not dissociated following DSB introduction, this synaptic complex could keep broken ends together for subsequent repair. In addition, the conservation of TA dinucleotides at IES ends, shown to be essential for DNA cleavage [13], may also contribute to precise repair. Indeed, all broken ends generated at IES boundaries exhibit the same characteristic geometry, with 4-base 5′ overhangs carrying a central TA that could guide their partial pairing.

IES excision starts after a few rounds of genome amplification have taken place in the developing MAC and 16 to 32 copies of each germline chromosome may be present when the first DSBs are introduced [16]. At least for the first excision events, we believe that NHEJ, which has long been referred to as error-prone, is the major pathway involved in the closure of excision sites. Homologous recombination is unlikely to account for these first events, since the new MAC differentiates from a copy of the germline nucleus, which only harbors the nonrearranged version of the genome: therefore, no rearranged template DNA is expected to preexist in the new MAC when it starts to develop. Thus, like V(D)J recombination in vertebrates, IES excision is a developmentally programmed DNA elimination relying on a domesticated transposase for the introduction of initiating DSBs and on the NHEJ pathway for the joining of flanking DNA ends. However, the two systems differ strikingly with regard to precision, V(D)J recombination joints being largely variable [4]. Our results support the notion that the diversity of the chromosomal junctions generated by V(D)J rearrangements does not result from inherent imprecision of the NHEJ pathway. An alternative explanation may reside in differences in the structure of the broken ends generated in the two systems or in their different processing. Rag1/Rag2-mediated cleavage generates blunt ends on the excised intervening fragment, which are ligated precisely to form a signal joint, while DNA hairpins are formed at flanking chromosome ends, which are opened in an imprecise manner, processed *via* nucleotide loss and/or addition before being joined by Lig4p-Xrcc4p. In contrast, cleaved ends generated in a PiggyMac-dependent manner during IES excision can readily align and pair, and require only limited processing before the final ligation step. *Paramecium*, therefore, provides a novel example of an essential role of NHEJ in the precise repair of developmentally programmed DSBs at a genome-wide scale.

Materials and Methods

Bioinformatics and phylogeny

LIG and *XRCC4* genes were identified in the macronuclear genome of *P. tetraurelia* by BLAST searches at ParameciumDB (http://paramecium.cgm.cnrs-gif.fr/) [45]. For DNA ligases, we first selected the genes that blasted against ligase genes from other organisms in the automated annotation of the MAC genome sequence [17]. In parallel, we used human ligases for tblastn search on the *Paramecium* genome. Finally, all putative *Paramecium* sequences were blasted against the whole genome to ensure that no paralog would be omitted and manual curation of gene annotations was carried out. As a query sequence for Xrcc4p, we used the 23-aa peptide from the human homolog, which was shown previously to interact with Ligase IV [46].

Phylogenetic trees were constructed using MEGA 4 Neighbor-Joining algorithm [47]. For Figure 1, we used a ClustalW alignment obtained from the NPS server [48].

P. tetraurelia strains and growth conditions

For autogamy time-course experiments, we used *P. tetraurelia* strain 51 (51 new in [11]) and its 51Δ*A* variant carrying a heritable deletion of the *A* gene in its MAC but harboring a wild-type MIC [13]. Cells were grown at 27°C in a wheat grass infusion (WGP; Pines International Inc.) inoculated with *Klebsiella pneumoniae* as described [16]. Autogamy was monitored by 4′-6-diamidino-2-phenylindole (DAPI)-staining. Total RNA and genomic DNA were extracted from ~400,000 cells for each time-point, as described [12]. Genomic DNA was quantified using the QuBit assay kit (Invitrogen) and RNA concentrations were estimated by absorption at 260 nm.

Derivatives of strain 51 were used for conjugation experiments. For genetic analysis, we used strain a3093 (mt7) homozygous for *pwB-96* and *nd9-c* and strain d4-502 (mt8) homozygous for *pwA-502* and *nd6-1* (kindly supplied by Mihoko Takahashi, University of Tsukuba). Cells were grown at 27°C in a pea medium inoculated with *K. pneumoniae* as described [34].

P. tetraurelia stock d4-110 (*hr-b/hr-b*) was used for synchronized conjugation: following mixing of reactive cells, conjugating pairs were synchronized and concentrated using iron dextran particles and strong neodymium magnets [34]. Total RNA was isolated from 50- to 100-mL aliquots of *Paramecium* cell culture (100 to 1,000 cells/mL), using the RNeasy Mini kit (QIAGEN) supplemented by a QIA shredder (QIAGEN) for homogenization and the RNase-free DNase set (QIAGEN) for genomic DNA elimination [34].

For localization of the Lig4p-GFP fusion, plasmid pLIG401GC was injected into the MAC of vegetative a3093. It contains the 496-bp genomic region upstream of the *LIG4a* open reading frame, the whole *LIG4a* open reading frame with the green fluorescent protein (GFP) gene fused to its 3′ end and the 3′UTR of gene G^{156} from *P. primaurelia* (sequence available upon request). Plasmid pRB35, containing the wild-type *pwB* gene with its 114-bp upstream and 233-bp downstream regions, was used to screen for successfully injected cells based on the complementation of the *pwB* phenotype and injected cells were mated with reactive d4-502 (*pwA*). A few hours after pair separation, two sub-populations could be sorted out according to their swimming phenotype when transferred to 20 mM KCl-containing medium. Exconjugants from injected cells exhibited a wild-type phenotype and swam backward for >30 secs, while those from *pwA* mutants swam backward slower and for shorter times. Details of GFP fluorescence analysis are provided in Text S1.

Molecular procedures

Oligonucleotides were purchased from Sigma-Aldrich or Eurofins MWG Operon (see Table S1).

PCR amplifications were performed in a final volume of 25 µL, with 10 pmol of each primer, 5 nmol of each dNTP and 1 U of DyNAzyme II DNA polymerase (Finnzymes), using an Eppendorf Mastercycler personal thermocycler. PCR products were analyzed on 3% NuSieve GTG agarose gels (BioWhittaker Molecular Applications). LMPCR detection of double-strand breaks was

performed as described [13]. Terminal transferase was used for poly(C) tailing of free 3′ ends and, following synthesis of the complementary strand from the Anchor(G) primer, tailed products were amplified by PCR as described [11]. All DNA sequencing was performed by GATC Biotech.

Northern blot and dot-blot hybridization were carried out as described in [12] for autogamy time-course experiments and in [34] for conjugation experiments.

RNA interference by feeding

RNAi plasmids. All RNAi plasmids are derivatives of vector L4440 [49] and carry a target gene fragment between two convergent T7 promoters (see Text S1 for a detailed description).

RNAi during conjugation. *P. tetraurelia* reactive cells (mt7 *pwB nd9* and mt8 *pwA nd6*) were obtained by feeding on *Escherichia coli* bacteria producing double-stranded RNA from plasmid pLIG4b-L [34]. Conjugation of RNAi-treated cells was induced within 48 hrs after inoculating the culture with *E. coli*. Conjugating pairs were transferred to standard *K. pneumoniae* medium and allowed to grow. For genetic analysis, F1 exconjugants from each pair were separated and grown for about 10 cell divisions before phenotypes were scored. Wild-type swimming behavior was indicative of successful conjugation, while a mutant phenotype in F1 revealed either conjugation failure or parental MAC regeneration. To distinguish between the last two possibilities, about 10 cells from each starved F1 were transferred to standard medium and allowed to grow before undergoing autogamy. F2 progeny of cells issued from a failed conjugation event should all exhibit a mutant phenotype, while 25% wild-type cells are expected in the post-autogamous progeny of MAC regenerants.

During conjugation, cells were fixed and treated with RNase and nuclear DNA was stained by VECTASHIELD with propidium iodide (see Text S1).

RNAi during autogamy. RNAi during autogamy was performed on strains 51 or 51Δ*A* as described [12]. At day 1 of starvation, cells were generally 100% autogamous and survival of their progeny was tested at day 2 or day 4 by transferring 30 individual autogamous cells to standard *K. pneumoniae* medium.

To monitor MAC regeneration, strain 51 was injected with *Bgl*I-restricted pZC'ΔRI, a modified pUC vector carrying the GFP-coding sequence adapted to *Paramecium* codon usage under the control of *P. primaurelia* G^{156} transcription signals (sequence available upon request). Following autogamy, GFP fluorescence was observed on living cells under a Leica fluorescence binocular. Presence of the GFP transgene in regenerated MACs was confirmed by dot-blot hybridization of total genomic DNA with a ^{32}P-labeled GFP probe.

Supporting Information

Figure S1 Phylogenetic tree of ATP-dependent DNA ligases. Amino-acid sequences of 61 ATP-dependent DNA Ligases were aligned using the MUSCLE algorithm [50]. Non-informative regions were removed manually from the alignment. The phylogenetic tree was generated by neighbor-joining using MEGA 4, with the following parameters: bootstrap 1000, pairwise deletion of gaps, equal input model, heterogeneous pattern among lineages, gamma distributed rates among site (with a gamma shape parameter = 1.7, as estimated by the means of gamma parameters calculated with the PhyML algorithm at http://www.hiv.lanl.gov). Bootstrap values are not shown when equal to 100. Prokaryotic NAD+-dependent replicative ligases (LigA) are not included. Some bacteria possess ATP-dependent DNA ligases (Ligases C or D) involved in DNA repair [51,52]. Some also encode ATP-dependent type K Ligase homologues: 3 randomly selected bacterial LigK are represented. ParameciumDB accession numbers of *Paramecium* genes are provided in the legend to Figure 1. TtL1a = XP_001022972.1 on GenBank. For all other proteins, Uniprot accession numbers are: TtL1b = Q24FD9, TtL4 = Q23RI5, TtlK = Q233G4, LmaL1 = Q4Q6U5, LmaKa = Q4Q960, LmaKb = Q4Q959, TbrL1 = Q587E4, TbrKa = Q6V9I8, TbrKb = Q56AN9, TcrL1a = Q4DX91, TcrKa = Q4DMH8, TcrKb = Q4DMH7, AtL1 = Q9C9M5, AtLIG1 = Q42572, AtL4 = Q9LL84, HsL4a = P49917, MmL1 = P37913, MmL4 = Q8BTF7, SpL = Q9C1W9, SpL4 = O74833, MkLb = Q8TWN3, MmeL1 = A6VFQ9, VvL = P33798, AfLD = O28549, AfLB = O29632, AtuLDa = A9CLR5, Cj = Q5HSC4, BpLD = Q63I59, HiL = P44121, MaLC = Q744K0, MaLD = Q742F5, MlLC = Q98NY5, MlLDa = Q98DP8, Nm = C6S4M8, PaLD = Q9I1X7, SwLD = Q0AXX1.
(TIF)

Figure S2 Conservation of the amino acids responsible for the Xrcc4–DNA ligase IV interaction. A multiple sequence alignment of the Lig4p-interacting region of Xrcc4p homologs from different organisms is displayed in the top panel. The linker regions connecting the BRCT domains of Lig4p homologs and involved in the interaction with Xrcc4p, are aligned in the bottom panel. Amino acids (aa) involved in the interaction between human ligase IV and Xrcc4p are shown (|). Strongly conserved residues are in cyan, identical in bold and red, and essential residues are highlighted in yellow [46]. Hs: *Homo sapiens*, At: *Arabidopsis thaliana*, Dm: *Drosophila melanogaster*, Sc: *Saccharomyces cerevisiae*, Pt: *Paramecium tetraurelia*.
(TIF)

Figure S3 Expression of *P. tetraurelia* DNA ligase and NHEJ genes during autogamy. NimbleGen whole-genome microarrays carry 6 oligonucleotide probes per gene: each slide was hybridized with a cDNA sample and the median of the six signals was calculated ([29], data available in ParameciumDB). Because all oligonucleotide probes present on the slides do not have the same Tm, microarrays do not allow comparing absolute transcript levels for different genes and only provide information on the relative variations of expression for each individual gene during autogamy. Following clusterization of slides, autogamy stages were defined as follows: VEG, vegetative cells; MEI, MIC meiosis; FRG, fragmented old MAC and no detectable new MAC following DAPI staining; DEV, visible developing new MACs (1, 2, 3 refer to three successive stages). Expression profiles are drawn from the mean values obtained for each stage from 4 independent time-course experiments. A. *Paramecium* ATP-dependent DNA ligases. For the particular case of *LIG4b*, automatic annotation of the draft MAC *P. tetraurelia* genome has split the gene into two open reading frames (GSPATG00022021001 and GSPATG00022020001). Thus, two sets of 6 probes contribute to the *LIG4b* signal and the curve represents the median of all 12 probes for each stage. This analysis reveals that two out of three *LIG1* genes are induced early during sexual processes, as expected for a postulated role in DNA replication, during meiosis or MAC development. The four *LIGK* are strongly upregulated at later stages during MAC development: future work should provide more insight into their function. B. Core NHEJ genes present in the *P. tetraurelia* genome. *XRCC4* accession number in ParameciumDB can be found in the legend to Figure 4A. Other ParameciumDB accession numbers are: GSPATG00006445001 for *KU70a*, GSPATG00009747001 for *KU70b*, GSPATG00034664001 for *KU80a*, GSPATG00035446001 for *KU80b* and GSPATG00030095001 for *KU80c*. Putative *CERNUNNOS* (*CER*) genes are also represented. Identi-

fication of *CER* genes (I. Callebaut, pers. comm.) was carried out by PSI-BLAST search and Hydrophobic Cluster Analysis (HCA) [53–55], as described in [56]. Both *CERa* and *CERb* were reannotated following manual curation: accession numbers in ParameciumDB are PTETG200014001 and PTETG2200011001, respectively.
(TIF)

Figure S4 Quantification of *LIG4* and *XRCC4* mRNA levels in silenced cells. Northern blots of total RNA extracted during different autogamy time-courses were probed with *LIG4* (A, B & C) or *XRCC4* (D) probes. All values were normalized with 17S rDNA signal for quantification. VK corresponds to vegetative cells grown in standard bacteria before transfer to silencing medium. The t = 0 hrs time-point was set as the time when 50% of cells had a fragmented old MAC. Times are in hours. A. RNAi was carried out on strain 51 (same experiment as in Figure 6), using pLIG4b-L (Figure 1) to induce the production of *LIG4* dsRNA. Total RNA samples extracted from both *LIG4*- and *ND7*-silenced cells were blotted onto a single membrane and hybridized to a *LIG4* probe. Time-points of *ND7* and *LIG4* RNAi experiments are displayed separately. B. RNAi against both *LIG4* genes was performed on strain 51Δ*A* (same experiment as in Figure 7 and Figure S6), using pLIG4a-R and pLIG4b-R (Figure 4A) for dsRNA production. Samples from *LIG4* and *ND7* silencing experiments were transferred separately to two different northern blots and the VK sample was loaded in duplicate on each membrane (1: *LIG4* blot; 2: *ND7* blot). The two blots were hybridized together with a *LIG4* probe. C. Three different dsRNA-producing constructs were tested separately to induce RNAi against *LIG4* genes in strain 51Δ*A*. All samples were loaded on the same blot and hybridized with a *LIG4* probe. Silencing using pLIG4a-R or pLIG4b-R was found to be as efficient as with pLIG4b-L, based on northern blot hybridization and on high lethality rates in the progeny of silenced cells. Autogamy stages are displayed below each diagram. V: vegetative cells; M: MIC meiosis; F: fragmented old MAC; NM: two visible new developing MACs; PA: post-autogamous cells with one MAC and surrounding fragments. D. RNAi against *XRCC4* was applied to strain 51Δ*A* using plasmid pXRCC4-L (Figure 4A). RNA samples from *XRCC4*- and *ND7*-silenced cells (same experiment as in Figure S7) were transferred to two separate membranes but hybridized together with an *XRCC4* probe. VK sample was loaded on both membranes (1: *LIG4* blot; 2: *ND7* blot). St = starved (St1N = -11 hrs; St2N = -4 hrs; St1X = -13 hrs; St2X = -1,5 hrs).
(TIF)

Figure S5 Normal progression of meiosis during the silencing of *LIG4* or *XRCC4*. DAPI staining of single cells fixed during autogamy in the macronuclear variant 51Δ*A*. Cells were not treated with RNase prior to staining. During *LIG4* silencing (A, C & D), starved cells were fixed 1 day following transfer to RNAi medium. For *XRCC4* silencing (B), the sample was treated at day 2. White arrows: micronuclear meiotic products. The heavily stained nucleus in each panel corresponds to the MAC. A. First meiotic division. B, C and D. After meiosis II. The eight haploid mics are sometimes not all visible, when they are not in the same focus.
(TIF)

Figure S6 LMPCR detection of DSBs during autogamy of the macronuclear variant 51Δ*A* silenced for *LIG4*. Four IESs are presented: 1 = 51A1835 (28 bp), 3 = 51A4404 (77 bp) from surface antigen *A* gene; 5 = sm19-576 (66 bp) from *SM19* tubulin gene and 6 = 51G4404 (222 bp) from surface antigen *G* gene. Gene names are indicated on the left of panels A and B. MAC flanking sequences are drawn as black lines and IESs as grey boxes. Vertical arrows indicate the position of DNA cleavage in each experiment. In all time-courses, cells were transferred to RNAi medium at day 0 (D0). The T0-time point is the time when 50% of cells have a fragmented old MAC. RNAi against *LIG4* was obtained using mixed bacterial cultures producing dsRNA from *LIG4a* (pLIG4a-R) and *LIG4b* (pLIG4b-R). V: vegetative cells. F: fragmented parental MAC. NM: cells with two visible new developing MACs. PA: post-autogamous cells with one MAC and surrounding fragments. Histograms on top show the progression of autogamy.
(TIF)

Figure S7 Molecular analysis of IES excision in *XRCC4*-silenced cells. Time-courses were performed with strain 51Δ*A* (same experiment as in Figure S4D). In each experiment, cells were transferred to RNAi medium at day 0 (D0) and starved to induce autogamy. V: vegetative cells. The T0-time point was arbitrarily chosen as the time when 50% of cells had a fragmented old MAC. F: cells with fragmented parental MAC. NM: cells with two new visible developing MACs. PA: post autogamous cells with only one new MAC and surrounding fragments. Only 3 IESs are shown: 3 = 51A4404 (77 bp), 5 = sm19-576 (66 bp) and 6 = 51G4404 (222 bp). Black lines: flanking MAC sequences, grey boxes: eliminated IES. A. Detection of excision junction for IES 51A4404. PCR around the IES allows the detection of a newly formed excision junction in the control but not in *XRCC4* silencing. Black arrowheads represent PCR primers. B. LMPCR analysis of double-strand breaks at IES boundaries. For each time-course, the progression of autogamy is displayed as a histogram of successive stages. Arrows indicate the position of DNA cleavage revealed by each molecular analysis.
(TIF)

Figure S8 Detection of single end-cleaved and linear excised IES molecules in *LIG4*-silenced cells. A & B. Detection of free 3′OH ends at the boundaries of IES #5 (sm19-576) during autogamy of strain 51Δ*A* silenced for *ND7* (Control) or *LIG4* expression. On each diagrammed molecule, the IES is represented by a grey box and its flanking DNA by a thick black line. The 3′ end tailed by the terminal transferase is indicated by a blue arrowhead. Poly(C)-tailed ends were amplified using primers sm19-5 and I (in A) or sm19-3 and I (in B, top). Ethidium bromide staining of 3% Nusieve agarose gels allowed only the detection of broken chromosome ends (A: expected size ~211 bp; B: expected size ~142 bp), while single-end cleaved IES molecules attached to their flanking DNA at their uncleaved end were not visible (expected sizes: ~277 bp in A and ~208 bp in B). Relevant size standards (in bp) are indicated on the left of each panel. In B (bottom), molecules carrying a DSB at the IES right end and still attached to their flanking DNA at the left end were revealed by ^{33}P-labeled primer extension (using nested primer sm19-3-aval), followed by electrophoresis on high resolution denaturing gels. The identity of PCR and primer extension products was confirmed by sequencing of gel-purified DNA. C. Sequencing chromatograms of the poly(C)-tailed chromosome ends generated by DNA cleavage at the left boundary of IES #5 (“MAC left” molecules in B). D. LMPCR-mediated detection of excised linear forms of IES #6 (51G4404) during autogamy of strain 51 silenced for *ND7* (control) or *LIG4* expression. Following the ligation of linker (ATAC)J’/I’, PCR amplification was carried out with primer I’ only. 51G4404-specific products were revealed by Southern blot hybridization of 3% Nusieve agarose gels, using the ^{32}P-labeled IES as a probe. The molecules to which the linker may be ligated are diagrammed below the pictures, with the linker and

primer I' represented by a purple box and arrowhead, respectively.
(TIF)

Table S1 Oligonucleotides used in this study. Black arrowheads show the position of each primer relative to a particular IES (drawn as a double line). (-) represents MAC flanking sequences. Acc. Nb: ParameciumDB accession number. GenBank: when necessary, the GenBank accession number corresponding to MIC sequences is displayed. O: orientation relative to the transcriptional direction of each gene. [bp]: length of the expected PCR product. T°C: annealing temperature used for PCR amplification. PCR conditions were as follows: 2 min at 95°C, 25–37 cycles of 20–60 sec at 95°C, 20–60 sec at the appropriate annealing temperature, and 1 min at 72°C, followed by a final step of 3 min at 72°C. A. Primers in *LIG4a* gene. B. Primers in *XRCC4* gene. Two product sizes are indicated (with/without additional nucleotides on the 5′ end of each primer, in lower-case). The two annealing temperatures that were used successively for PCR amplification are shown. C. Primer in the 17S rDNA gene. D. Primers for MAC junction analysis. The two product lengths correspond to IES+ or IES- forms, respectively. E. Primers for circle junction analysis. For IES 51A2591, the two products lengths correspond to forms with or without the 28-bp internal IES, respectively. F. Primers used for LMPCR to detect free broken chromosome ends. Red arrowheads mark the position of the ligated linker and of primer PCRhaut, which hybridizes within the linker. PCRhaut is used to amplify ligation products, in combination with a second primer represented by a black arrowhead. Oligonucleotides used for primer extension (1 to 10 cycles) are drawn as white arrowheads. G. Primers used for LMPCR to detect free broken IES ends. Diagrams are as in F, with I' instead of PCRhaut. H. Primers used for the detection of free 3′OH ends at MAC ends. After a first elongation step with Anchor(G) (which hybridizes to the poly(C) tail added by TdT), the DNA is amplified by PCR using I and specific primers.
(PDF)

Text S1 Supplemental experimental procedures.
(PDF)

Acknowledgments

We thank Céline Baudry, Chloé Hot, and students of the 2007–2008 Pasteur Course on "Analysis of Genomes" for contributing to RNAi experiments; Olivier Arnaiz and Linda Sperling for their support in bioinformatic analyses; and Jacek Nowak, Julien Bischerour, Vinciane Régnier, and Bénédicte Michel for stimulating discussions. We are grateful to Isabelle Callebaut for her help in identifying the *Cernunnos* homologs. Many thanks to Linda Sperling for critical reading of the manuscript.

Author Contributions

Conceived and designed the experiments: A Kaputsta, A Matsuda, JD Forney, S Malinsky, M Bétermier. Performed the experiments: A Kapusta, A Matsuda, A Marmignon, M Ku, A Silve, S Malinsky, M Bétermier. Analyzed the data: A Kapusta, A Matsuda, A Marmignon, E Meyer, JD Forney, S Malinsky, M Bétermier. Contributed reagents/materials/analysis tools: E Meyer, JD Forney, M Bétermier. Wrote the paper: A Kapusta, S Malinsky, M Bétermier.

References

1. Keeney S, Neale MJ (2006) Initiation of meiotic recombination by formation of DNA double-strand breaks: mechanism and regulation. Biochem Soc Trans 34: 523–525.
2. Soulas-Sprauel P, Rivera-Munoz P, Malivert L, Le Guyader G, Abramowski V, et al. (2007) V(D)J and immunoglobulin class switch recombinations: a paradigm to study the regulation of DNA end-joining. Oncogene 26: 7780–7791.
3. Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 3: e181. doi:10.1371/journal.pbio.0030181.
4. Gellert M (2002) V(D)J recombination: RAG proteins, repair factors, and regulation. Annu Rev Biochem 71: 101–132.
5. Lieber MR (2010) The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annu Rev Biochem.
6. Jahn CL, Klobutcher LA (2002) Genome remodeling in ciliated protozoa. Annu Rev Microbiol 56: 489–520.
7. Yao MC, Duharcourt S, Chalker DL (2002) Genome-wide rearrangements of DNA in ciliates. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, eds. Mobile DNA II. Washington, D.C.: ASM Press. pp 730–758.
8. Prescott DM (1994) The DNA of ciliated protozoa. Microbiol Rev 58: 233–267.
9. Le Mouel A, Butler A, Caron F, Meyer E (2003) Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in paramecia. Eukaryot Cell 2: 1076–1090.
10. Bétermier M (2004) Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate Paramecium. Res Microbiol 155: 399–408.
11. Gratias A, Bétermier M (2003) Processing of double-strand breaks is involved in the precise excision of paramecium internal eliminated sequences. Mol Cell Biol 23: 7152–7162.
12. Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, et al. (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev 23: 2478–2483.
13. Gratias A, Lepere G, Garnier O, Rosa S, Duharcourt S, et al. (2008) Developmentally programmed DNA splicing in Paramecium reveals short-distance crosstalk between DNA cleavage sites. Nucleic Acids Res 36: 3244–3251.
14. Saveliev SV, Cox MM (1995) Transient DNA breaks associated with programmed genomic deletion events in conjugating cells of Tetrahymena thermophila. Genes Dev 9: 248–255.
15. Williams K, Doak TG, Herrick G (1993) Developmental precise excision of Oxytricha trifallax telomere-bearing elements and formation of circles closed by a copy of the flanking target duplication. Embo J 12: 4593–4601.
16. Bétermier M, Duharcourt S, Seitz H, Meyer E (2000) Timing of developmentally programmed excision and circularization of Paramecium internal eliminated sequences. Mol Cell Biol 20: 1553–1561.
17. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444: 171–178.
18. Ellenberger T, Tomkinson AE (2008) Eukaryotic DNA ligases: structural and functional insights. Annu Rev Biochem 77: 313–338.
19. Sinha KM, Hines JC, Downey N, Ray DS (2004) Mitochondrial DNA ligase in Crithidia fasciculata. Proc Natl Acad Sci U S A 101: 4361–4366.
20. Downey N, Hines JC, Sinha KM, Ray DS (2005) Mitochondrial DNA ligases of Trypanosoma brucei. Eukaryot Cell 4: 765–774.
21. McVey M, Lee SE (2008) MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet 24: 529–538.
22. Burton P, McBride DJ, Wilkes JM, Barry JD, McCulloch R (2007) Ku heterodimer-independent end joining in Trypanosoma brucei cell extracts relies upon sequence microhomology. Eukaryot Cell 6: 1773–1781.
23. Critchlow SE, Bowater RP, Jackson SP (1997) Mammalian DNA double-strand break repair protein XRCC4 interacts with DNA ligase IV. Curr Biol 7: 588–598.
24. Ahnesorg P, Smith P, Jackson SP (2006) XLF interacts with the XRCC4-DNA ligase IV complex to promote DNA nonhomologous end-joining. Cell 124: 301–313.
25. Buck D, Malivert L, de Chasseval R, Barraud A, Fondaneche MC, et al. (2006) Cernunnos, a novel nonhomologous end-joining factor, is mutated in human immunodeficiency with microcephaly. Cell 124: 287–299.
26. Frank KM, Sekiguchi JM, Seidl KJ, Swat W, Rathbun GA, et al. (1998) Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. Nature 396: 173–177.
27. Barnes DE, Stamp G, Rosewell I, Denzel A, Lindahl T (1998) Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. Curr Biol 8: 1395–1398.
28. Gao Y, Sun Y, Frank KM, Dikkes P, Fujiwara Y, et al. (1998) A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis. Cell 95: 891–902.
29. Arnaiz O, Gout JF, Betermier M, Bouhouche K, Cohen J, et al. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. BMC Genomics 11: 547.
30. Berger JD (1973) Nuclear differentiation and nucleic acid synthesis in well-fed exconjugants of Paramecium aurelia. Chromosoma 42: 247–268.
31. Kung C, Chang SY, Satow Y, Houten JV, Hansma H (1975) Genetic dissection of behavior in paramecium. Science 188: 898–904.

32. Galvani A, Sperling L (2002) RNA interference by feeding in Paramecium. Trends Genet 18: 11–12.
33. Berger JD (1973) Selective inhibition of DNA synthesis in macronuclear fragments in Paramecium aurelia exconjugants and its reversal during macronuclear regeneration. Chromosoma 44: 33–48.
34. Matsuda A, Forney JD (2006) The SUMO pathway is developmentally regulated and required for programmed DNA elimination in Paramecium tetraurelia. Eukaryot Cell 5: 806–815.
35. Berger JD (1974) Selective autolysis of nuclei as a source of DNA precursors in Paramecium aurelia exconjugants. J Protozool 21: 145–152.
36. Duharcourt S, Lepere G, Meyer E (2009) Developmental genome rearrangements in ciliates: a natural genomic subtraction mediated by non-coding transcripts. Trends Genet 25: 344–350.
37. Nowacki M, Landweber LF (2009) Epigenetic inheritance in ciliates. Curr Opin Microbiol 12: 638–643.
38. Matsuda A, Shieh AW, Chalker DL, Forney JD (2010) The conjugation-specific Die5 protein is required for development of the somatic nucleus in both Paramecium and Tetrahymena. Eukaryot Cell 9: 1087–1099.
39. Cheng CY, Vogt A, Mochizuki K, Yao MC (2010) A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in Tetrahymena thermophila. Mol Biol Cell 21: 1753–1762.
40. Nowacki M, Higgins BP, Maquilan GM, Swart EC, Doak TG, et al. (2009) A functional role for transposases in a large eukaryotic genome. Science 324: 935–938.
41. Garnier O, Serrano V, Duharcourt S, Meyer E (2004) RNA-mediated programming of developmental genome rearrangements in Paramecium tetraurelia. Mol Cell Biol 24: 7370–7379.
42. Budman J, Kim SA, Chu G (2007) Processing of DNA for nonhomologous end-joining is controlled by kinase activity and XRCC4/ligase IV. J Biol Chem 282: 11950–11959.
43. Mitra R, Fain-Thornton J, Craig NL (2008) piggyBac can bypass DNA synthesis during cut and paste transposition. Embo J 27: 1097–1109.
44. Guirouilh-Barbat J, Huck S, Bertrand P, Pirzio L, Desmaze C, et al. (2004) Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells. Mol Cell 14: 611–623.
45. Arnaiz O, Cain S, Cohen J, Sperling L (2007) ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data. Nucleic Acids Res 35: D439–444.
46. Sibanda BL, Critchlow SE, Begun J, Pei XY, Jackson SP, et al. (2001) Crystal structure of an Xrcc4-DNA ligase IV complex. Nat Struct Biol 8: 1015–1019.
47. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
48. Combet C, Blanchet C, Geourjon C, Deleage G (2000) NPS@: network protein sequence analysis. Trends Biochem Sci 25: 147–150.
49. Timmons L, Fire A (1998) Specific interference by ingested dsRNA. Nature 395: 854.
50. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
51. Gong C, Bongiorno P, Martins A, Stephanou NC, Zhu H, et al. (2005) Mechanism of nonhomologous end-joining in mycobacteria: a low-fidelity repair system driven by Ku, ligase D and ligase C. Nat Struct Mol Biol 12: 304–312.
52. Zhu H, Shuman S (2007) Characterization of Agrobacterium tumefaciens DNA ligases C and D. Nucleic Acids Res 35: 3631–3645.
53. Gaboriaud C, Bissery V, Benchetrit T, Mornon JP (1987) Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett 224: 149–155.
54. Callebaut I, Courvalin JC, Worman HJ, Mornon JP (1997) Hydrophobic cluster analysis reveals a third chromodomain in the Tetrahymena Pdd1p protein of the chromo superfamily. Biochem Biophys Res Commun 235: 103–107.
55. Eudes R, Le Tuan K, Delettre J, Mornon JP, Callebaut I (2007) A generalized analysis of hydrophobic and loop clusters within globular protein sequences. BMC Struct Biol 7: 2.
56. Callebaut I, Malivert L, Fischer A, Mornon JP, Revy P, et al. (2006) Cernunnos interacts with the XRCC4 x DNA-ligase IV complex and is homologous to the yeast nonhomologous end-joining factor Nej1. J Biol Chem 281: 13857–13860.

Figure S1

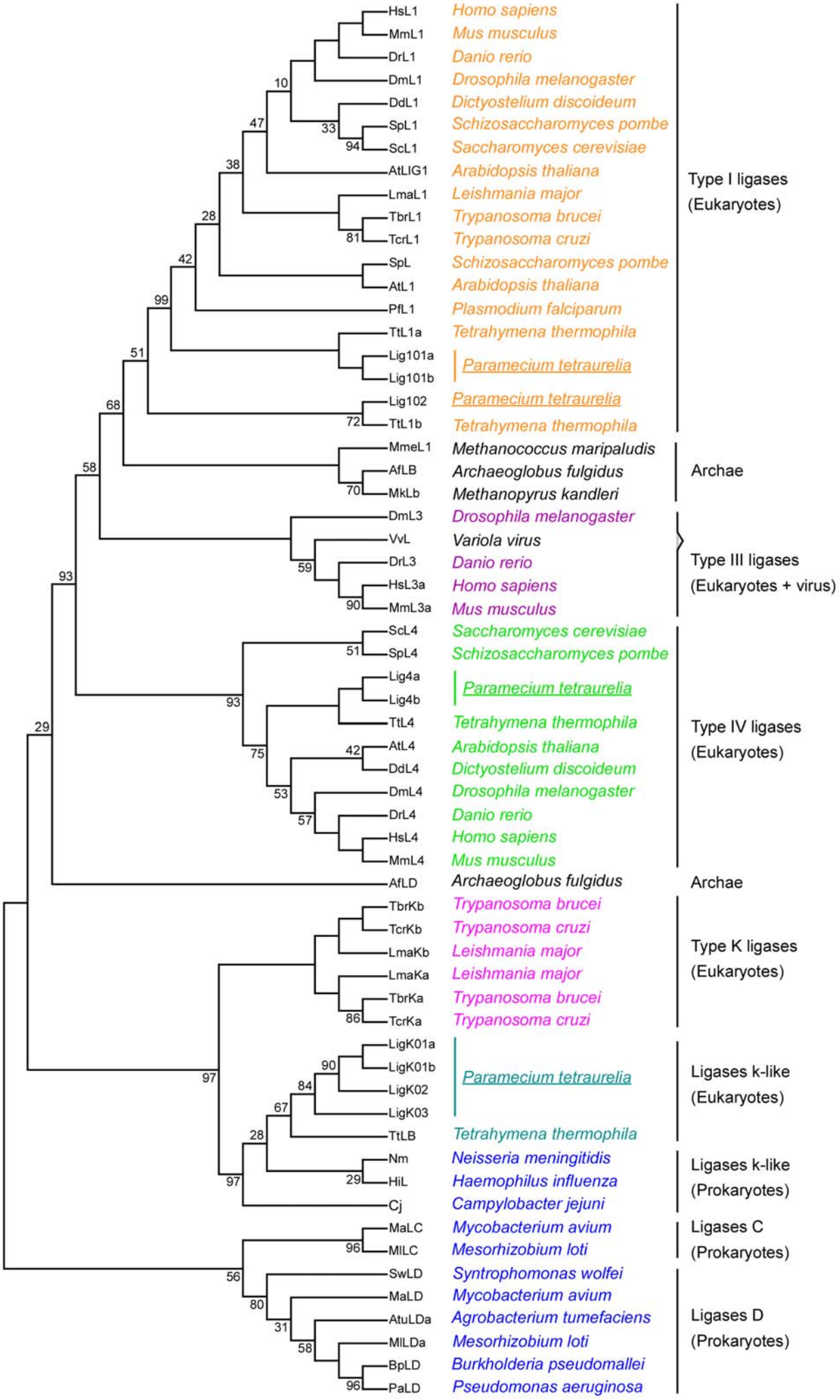

HsL1
MmL1
DrL1
DmL1
DdL1
SpL1
ScL1
AtLIG1
LmaL1
TbrL1
TcrL1
SpL
AtL1
PfL1
TtL1a
Lig101a
Lig101b
Lig102
TtL1b
MmeL1
AfLB
MkLb
DmL3
VvL
DrL3
HsL3a
MmL3a
ScL4
SpL4
Lig4a
Lig4b
TtL4
AtL4
DdL4
DmL4
DrL4
HsL4
MmL4
AfLD
TbrKb
TcrKb
LmaKb
LmaKa
TbrKa
TcrKa
LigK01a
LigK01b
LigK02
LigK03
TtLB
Nm
HiL
Cj
MaLC
MlLC
SwLD
MaLD
AtuLDa
MlLDa
BpLD
PaLD
Homo sapiens
Mus musculus
Danio rerio
Drosophila melanogaster
Dictyostelium discoideum
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Arabidopsis thaliana
Leishmania major
Trypanosoma brucei
Trypanosoma cruzi
Schizosaccharomyces pombe
Arabidopsis thaliana
Plasmodium falciparum
Tetrahymena thermophila
Paramecium tetraurelia
Paramecium tetraurelia
Tetrahymena thermophila
Methanococcus maripaludis
Archaeoglobus fulgidus
Methanopyrus kandleri
Drosophila melanogaster
Variola virus
Danio rerio
Homo sapiens
Mus musculus
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Paramecium tetraurelia
Tetrahymena thermophila
Arabidopsis thaliana
Dictyostelium discoideum
Drosophila melanogaster
Danio rerio
Homo sapiens
Mus musculus
Archaeoglobus fulgidus
Trypanosoma brucei
Trypanosoma cruzi
Leishmania major
Leishmania major
Trypanosoma brucei
Trypanosoma cruzi
Paramecium tetraurelia
Tetrahymena thermophila
Neisseria meningitidis
Haemophilus influenza
Campylobacter jejuni
Mycobacterium avium
Mesorhizobium loti
Syntrophomonas wolfei
Mycobacterium avium
Agrobacterium tumefaciens
Mesorhizobium loti
Burkholderia pseudomallei
Pseudomonas aeruginosa
Type I ligases (Eukaryotes)
Archae
Type III ligases (Eukaryotes + virus)
Type IV ligases (Eukaryotes)
Archae
Type K ligases (Eukaryotes)
Ligases k-like (Eukaryotes)
Ligases k-like (Prokaryotes)
Ligases C (Prokaryotes)
Ligases D (Prokaryotes)

Figure S2

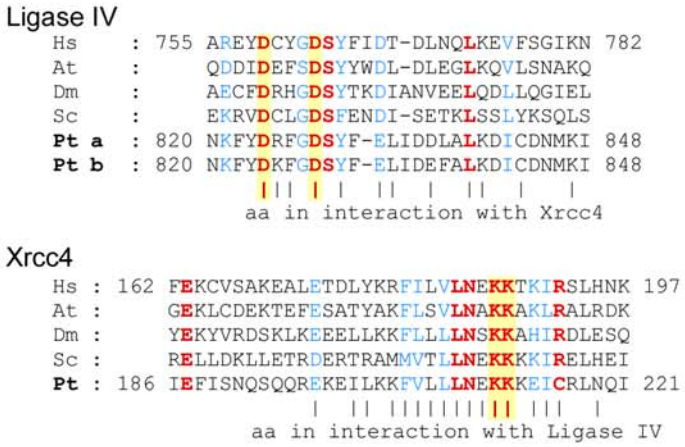

Ligase IV
Hs : 755 AREYDCYGDSYFIDT-DLNQLKEVFSGIKN 782
At : QDDIDEFSDSYYWDL-DLEGLKQVLSNAKQ
Dm : AECFDRHGDSYTKDIANVEELQDLLQGIEL
Sc : EKRVDCLGDSFENDI-SETKLSSLYKSQLS
Pt a : 820 NKFYDRFGDSYF-ELIDDLALKDICDNMKI 848
Pt b : 820 NKFYDKFGDSYF-ELIDEFALKDICDNMKI 848
aa in interaction with Xrcc4
Xrcc4
Hs : 162 FEKCVSAKEALETDLYKRFILVLNEKKTKIRSLHNK 197
At : GEKLCDEKTEFESATYAKFLSVLNAKKAKLRALRDK
Dm : YEKYVRDSKLKEELLKKFLLLLNSKKAHIRDLESQ
Sc : RELLDKLLETRDERTRAMMVTLLNEKKKKIRELHEI
Pt : 186 IEFISNQSQQREKEILKKFVLLLNEKKKEICRLNQI 221
aa in interaction with Ligase IV

Figure S3

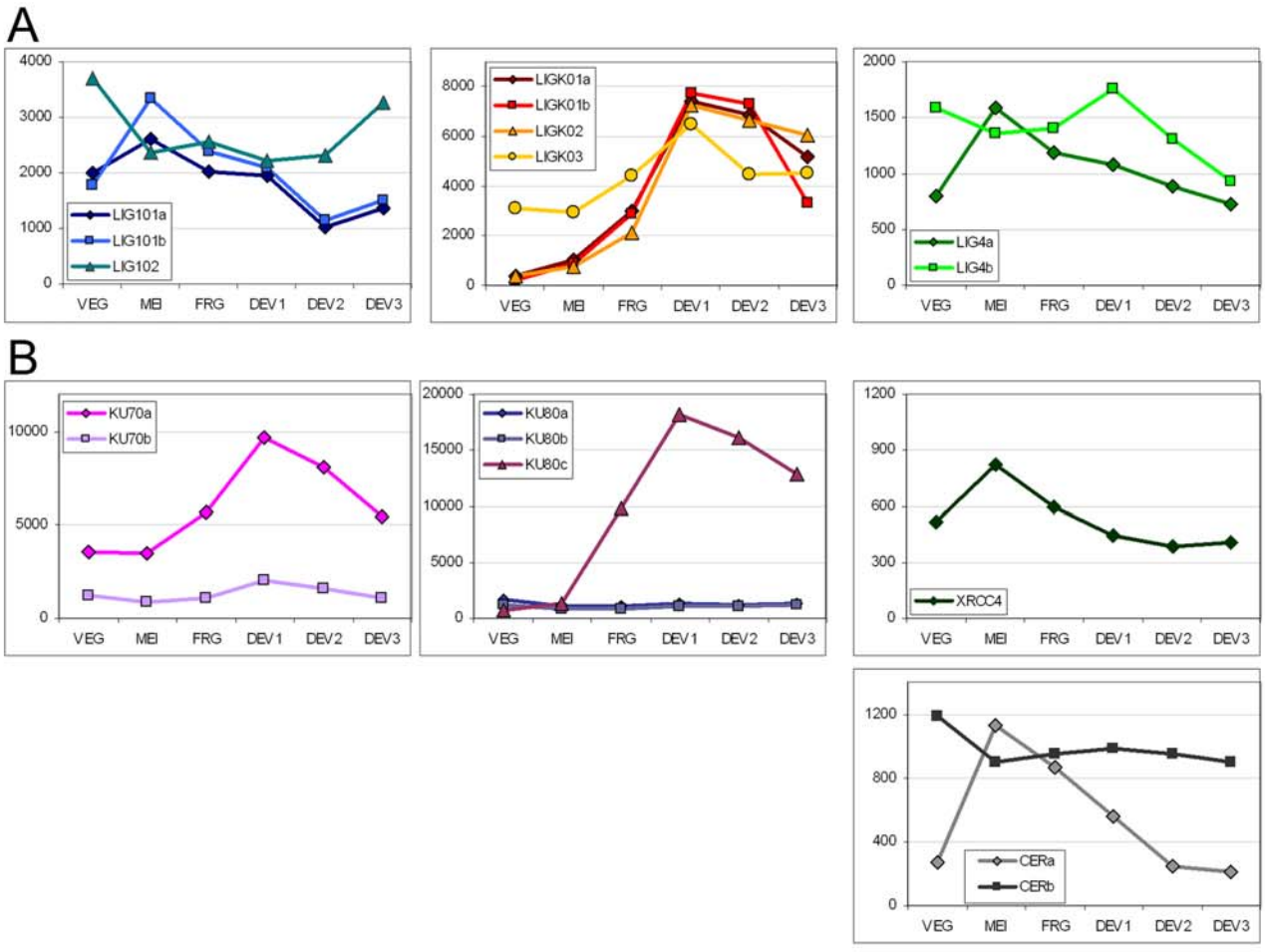

A
4000
3000
2000
1000
0
LIG101a
LIG101b
LIG102
VEG MEI FRG DEV1 DEV2 DEV3
8000
6000
4000
2000
0
LIGK01a
LIGK01b
LIGK02
LIGK03
VEG MEI FRG DEV1 DEV2 DEV3
2000
1500
1000
500
0
LIG4a
LIG4b
VEG MEI FRG DEV1 DEV2 DEV3
B
10000
5000
0
KU70a
KU70b
VEG MEI FRG DEV1 DEV2 DEV3
20000
15000
10000
5000
0
KU80a
KU80b
KU80c
VEG MEI FRG DEV1 DEV2 DEV3
1200
900
600
300
0
XRCC4
VEG MEI FRG DEV1 DEV2 DEV3
1200
800
400
0
CERa
CERb
VEG MEI FRG DEV1 DEV2 DEV3

Figure S4

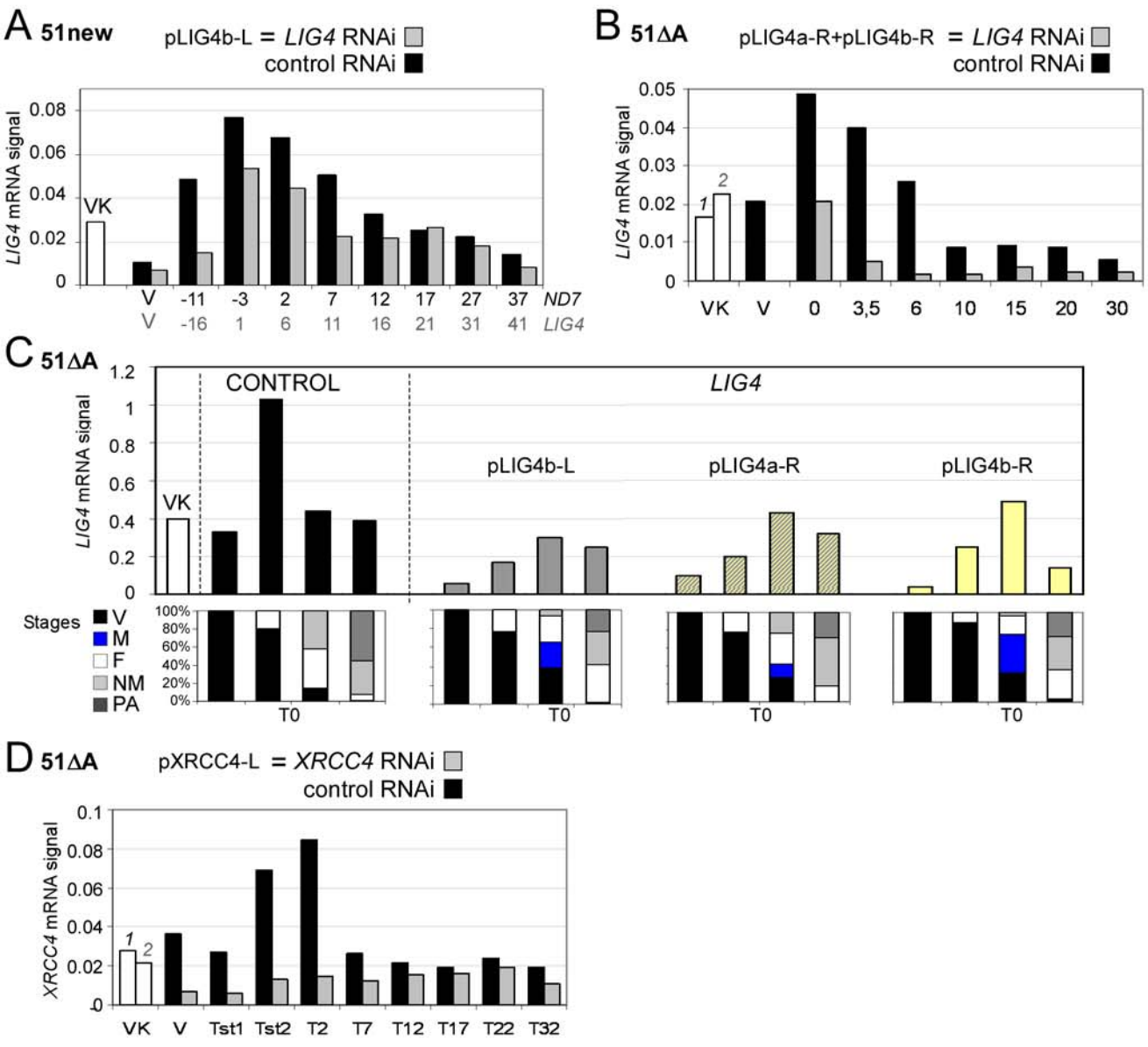

A 51new
pLIG4b-L = LIG4 RNAi
control RNAi
LIG4 mRNA signal
VK
V -11 -3 2 7 12 17 27 37 ND7
V -16 1 6 11 16 21 31 41 LIG4
B 51ΔA
pLIG4a-R+pLIG4b-R = LIG4 RNAi
control RNAi
LIG4 mRNA signal
VK V 0 3,5 6 10 15 20 30
C 51ΔA
CONTROL
LIG4
pLIG4b-L
pLIG4a-R
pLIG4b-R
LIG4 mRNA signal
Stages V M F NM PA
T0
D 51ΔA
pXRCC4-L = XRCC4 RNAi
control RNAi
XRCC4 mRNA signal
VK V Tst1 Tst2 T2 T7 T12 T17 T22 T32

Figure S5

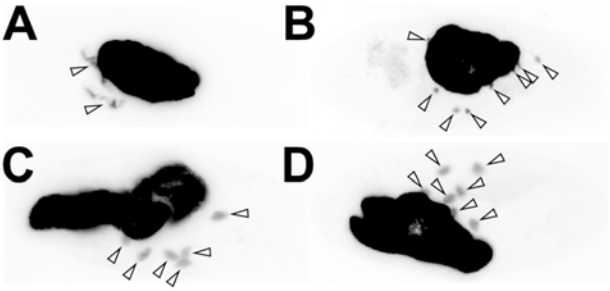


Figure S6

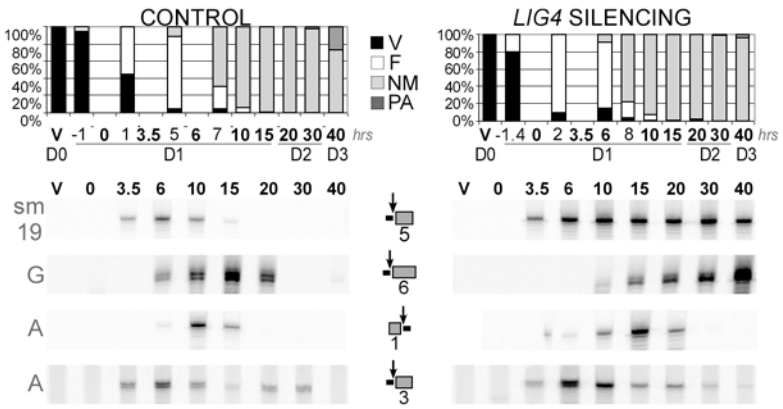

CONTROL
LIG4 SILENCING
100%
80%
60%
40%
20%
0%
V
F
NM
PA
V -1 0 1 3.5 5 6 7 10 15 20 30 40 hrs
V -1.4 0 2 3.5 6 8 10 15 20 30 40 hrs
D0
D1
D2
D3
V 0 3.5 6 10 15 20 30 40
sm 19
G
A
A
5
6
1
3

Figure S7

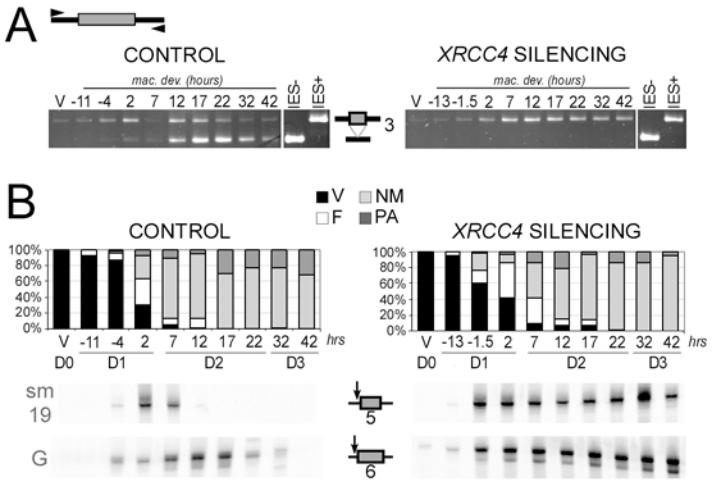
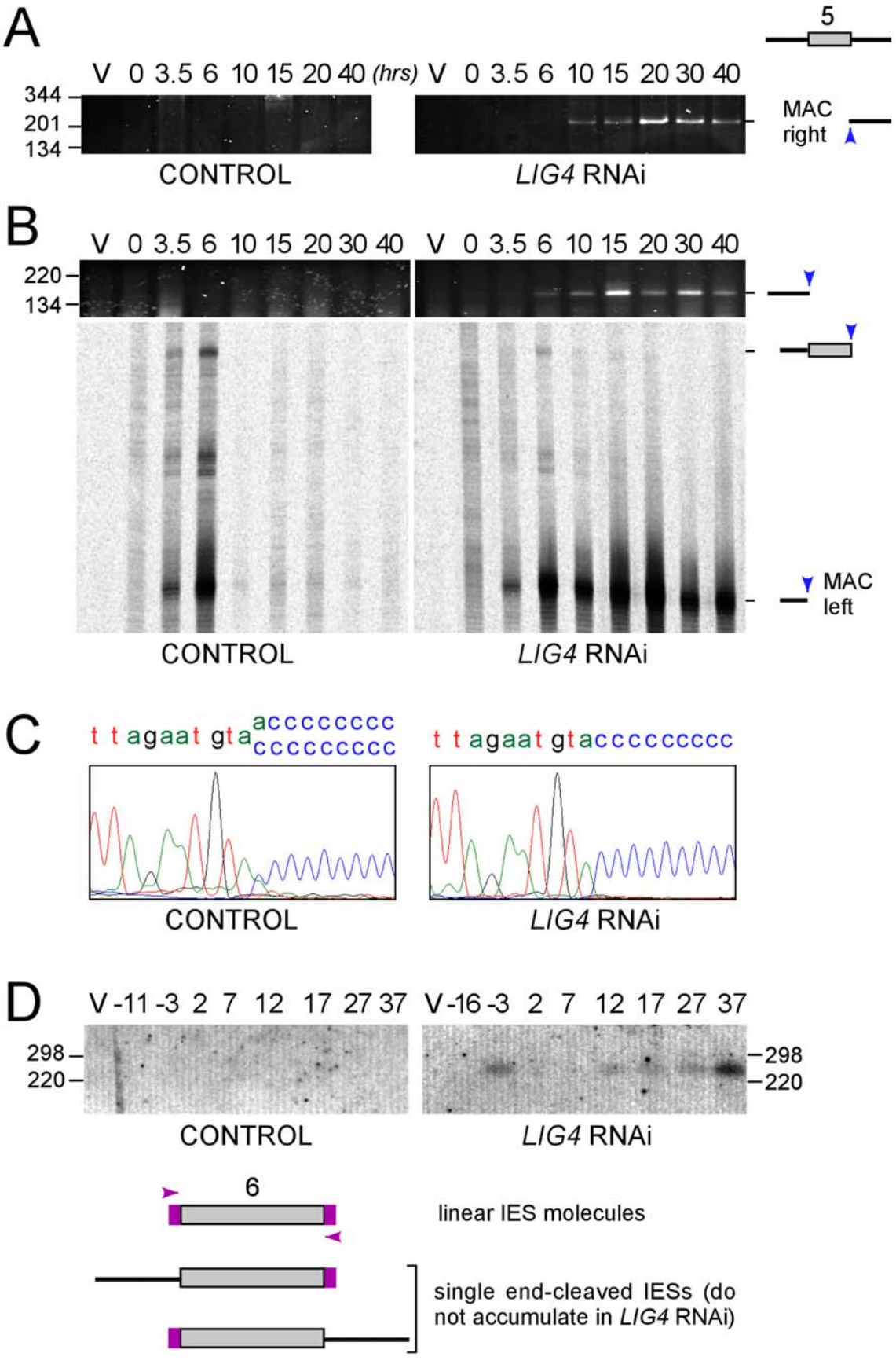

A
CONTROL
XRCC4 SILENCING
mac. dev. (hours)
V -11 -4 2 7 12 17 22 32 42
IES-
IES+
3
V -13 -1.5 2 7 12 17 22 32 42
B
V
NM
F
PA
100%
80%
60%
40%
20%
0%
V -11 -4 2 7 12 17 22 32 42 hrs
D0 D1 D2 D3
V -13 -1.5 2 7 12 17 22 32 42 hrs
sm
19
G
5
6

Figure S8


A
5
V 0 3.5 6 10 15 20 40 (hrs) V 0 3.5 6 10 15 20 30 40
344
201
134
MAC
right
CONTROL
LIG4 RNAi
B
V 0 3.5 6 10 15 20 30 40 V 0 3.5 6 10 15 20 30 40
220
134
MAC
left
CONTROL
LIG4 RNAi
C
t t agaat gta accccccccc
cccccccccc
t t agaat gtaccccccccc
CONTROL
LIG4 RNAi
D
V -11 -3 2 7 12 17 27 37 V -16 -3 2 7 12 17 27 37
298
220
298
220
CONTROL
LIG4 RNAi
6
linear IES molecules
single end-cleaved IESs (do
not accumulate in LIG4 RNAi)

A. Primers in *LIG4a* gene

Oligo name	Sequence (5'-3')	Note	Acc. Nb	O.	[bp]	T°C
ligIV-1_ATGup	ATGCAGCAGGAGAATTTTATAGACGAAGAG	Used for *LIG4a* probe	PTETG5400008001	+	510	62
ligIV-1_500R	ATTATGCACAACCATGGGGAAATCACCAAC			-		

B. Primers in *XRCC4* gene

Oligo name	Sequence (5'-3')	Note	Acc. Nb	O.	[bp]	T°C
XRCC4_302_BglII_U	atgcagatctAGGAAATAATGTATAAGAGTTTAGAAG	Used for *XRCC4* probe, in *XRCC4* silencing	GSPATG00029540001	+	322/	53/
XRCC4_302_PstI_R	atgcctgcagTTACTCTTCATATTGTTTCTTAACTGG			-	302	61.5
XRCC4-496_U	GAATAAAAAAGAAATTAAACAAGATGGCG	Used for probe in wild type, and as RNAi fragment	GSPATG00029540001	+	496	56
XRCC4-496_R	AAATCTCCTTGCAAACCCTGTAAAAATGG			-		

C. Primer in 17S rDNA

Oligo name	Sequence (5'-3')	Note
17Sext_NcoI	ACCCGTGACTGCCATGGTAGTCCAATACA	Used as oligo probe on Northern Blot

D. MAC junction analysis

IES	Oligo name	Sequence (5'-3')	Diagram	GenBank	O.	[bp]	T°C
A1835	51A1835-5'	TAATGTATTGATAAGGCTTGCTCTACAGCC	− ▶ − === − − −	L26124	+	240/211	60
	51A1835-3'(4)	AGACAAGTAGGGAATCCACTTCTAGTAATC	− − − === − ◀ −		-		
A2591	51A2591-5'	ACACCAAGCGAAACATGCACAGTCG	− ▶ − === − − −	L26124	+	476/106	54
	51A2591-3'	TTTTATGGCATTAAGCTTGTGTCAT	− − − === − ◀ −		-		
A4404	51A4404-5'	TAAATGTTCAGCTTACAACGCAGCT	− ▶ − === − − −	L26124	+	263/189	56
	51A4404-3'(2)	CCAGTTATTGAACTGCAACTTACTGCAGTG	− − − === − ◀ −		-		

E. Circle junction analysis

IES	Oligo name	Sequence (5'-3')	Diagram	GenBank	O.	[bp]	T°C
A2591	51A2591-09	CAATATTATACATCTAGAACTTATAGTTAG	− − − = ◀ = − − −	L26124	+	370/342	55.5
	51A2591-01	AGATTTATATCTTTTTTCTCAAATTCAGC	− − − = ▶ = − − −		-		
A4578	51A4578-2	TGGTTGTTAGTCTCAAAGAATTCTAAAGAC	− − − = ◀ = − − −	L26124	+	408	55.8
	51A4578-7	AAGAAATTTTATTGTAAATATATTTTCAGC	− − − = ▶ = − − −		-		
G4404	51G07	TTTTGAAATATTTTCAAGTTTTTGGACTAC	− − − = ◀ = − − −	AJ010441	+	222	54
	51G08	ACAATATATATTTACTTGATAATATTTTCC	− − − = ▶ = − − −		-		

F. LMPCR - Detection of free broken chromosome ends

IES	Oligo name	Sequence (5'-3')	Diagram	GenBank	O.	[bp]	T°C
	PCRhaut (►)	GAATTCGGATCCGCTCGGACCGTGGC	used with specific primer ► to amplify the ligated product				
A1835	51A1835-3'(3)	GTAGTACAAGATTTTTCGACACAAGTTGAG	– – – === ► – – ◄ –	L26124	-	225	60
	51A1835-3'(4)	AGACAAGTAGGGAATCCACTTCTAGTAATC	– – – === – – ◁ – –		-	180	62
	51A1835-5'(3)	GGTTGCGTAACACTTCCTCTTAAATGTGAG	– ► – – ◄ === – – –		+	206	63
	51A1835-5'(4)	GAAGTCTAATGGATAACCTTGTGGATGGAC	– – ▷ – – === – – –		+	147	63
A2591	51A2591-16	AATTGTAAATTGACTTCAGCAAATAAAAAA	– ► – – ◄ === – – –	L26124	+	218	60
	51A2591-17	ATGTGTTTGGACTGGATTGGCATGTAGAAG	– – ▷ – – === – – –		+	189	63
A4404	51A4404-5'(2)	TGGAATAGTGCTGCATCACCAGCTGCTTGC	– ► – – ◄ === – – –	L26124	+	177	63
	51A4404-5'(3)	ACCAGCTGCTTGCATTCAAATATCCACAGT	– – ▷ – – === – – –		+	160	63
G4404	51G18	ACTGTTGCTACACATTGTGCATATGTTACT	– ► – – ◄ === – – –	AJ010441	+	135	60
	51G13	TGCATATGTTACTGGAACTGGATTGGTAGC	– – ▷ – – === – – –		+	118	63
sm19-576	sm19-2	AATTAAGCAAGAAAAGAAATAGAAAAAACC	– ► – – ◄ === – – –	AJ272425	+	188	54
	sm19-3	CTACAATAATGAGTCTAGCTGGTGGCACTG	– – ▷ – – === – – –	(mac sequence)	+	139	63

G. LMPCR - Detection of free broken IES ends

IES	Oligo name	Sequence (5'-3')	Diagram	GenBank	O.	[bp]	T°C
	I' (►)	GCTCGGACCGTGGCTAGCATTAGTC	used with specific primer ► to amplify the ligated product				
G4404	51G11	GGACTACTTTTGAAATTGAATTATAACAAAGGC	– – ► === ◄ = – – –	AJ010441	-	154	60
	51G10	AAAGGCTAATTTGGATGAATGAGCATTAAATC	– – – == ◁ == – – –		-	127	60

H. Detection of free broken 3'OH ends

IES	Oligo name	Sequence (5'-3')	Diagrams and Notes		GenBank	O.	T°C
	I' (►)	GCTCGGACCGTGGCTAGCATTAGTC					
	Anchor(G)	GCTCGGACCGTGGCTAGCATTAGTGAGTGGGGGGGGGGGGGGG					
G4404	51G18	ACTGTTGCTACACATTGTGCATATGTTACT	– ► – – ◄ === – – –		AJ010441	+	60
	51G13	TGCATATGTTACTGGAACTGGATTGGTAGC	– – ▷ – – === – – –	(primer extension and sequencing)		+	63
sm19-576	sm19-3	CTACAATAATGAGTCTAGCTGGTGGCACTG	– ► – – ◄ === – – –	– ► – – === ◄ – – –	AJ272425 (mac)	-	64
	sm19-5	ATCCACACTTTTATCGATTTGCTTTGATCC	– – – === ► – – ◄ –	– – – ► === – – ◄ –		-	60
	sm19-3-aval	GTGGCACTGGATCCGGTCTAGGTAGTAG	– – – === – – ◁ – –	(primer extension and sequencing)		-	64
	sm19-6	ACTTAACATTTAATATATCCTGTCATTCTC	to sequence sm19-5/I PCR productss			-	

Supplemental experimental procedures

Analysis of GFP fusions

Approximately 2 pL of circular pLIG401GC (3 μg/μL) and pRB35 (1 μg/mL) in distilled water were injected into the macronucleus of vegetative a3093 cells under an inverted light microscope. For fixation of GFP-expressing cells, about 10-20 cells were concentrated by removing surrounding medium to less than 10 μL, then quickly mixed with 4% paraformaldehyde in phosphate-buffered saline, pH 7.4 (PBS). Cells were washed twice with PBS, treated with 0.1 mg/ml RNaseA, then mounted and counterstained by VECTASHIELD with DAPI (Vector Laboratories, Burlingame, CA). For quantitative observations, cells were flattened by gentle pressure on coverslips, with $\sim$5 μL rice glue cushions at the four corners. Fluorescence microscope (Olympus, Japan) images were obtained using a CCD camera (Spot, France) with fixed gain and exposure time.

Quantitative measurements of GFP signals were made with ImageJ (National Institutes of Health) and total fluorescence was calculated as follows:

$$\text{Total fluorescence (count)} = (I - B) \times A$$

where I is the mean pixel intensity measured for each object (mating pairs or individual starved or exconjugant cells), B is the background pixel intensity and A (in pixels) is the area of the object (measured under visible light).

RNAi plasmids

All RNAi plasmids are derivatives of vector L4440 (Timmons and Fire, 1998) and carry a target gene fragment between two convergent T7 promoters. Cloned fragments were as follows: nt 521-1774 from *LIG4b* for pLIG4b-L, 1824-2596 from *LIG4a* for pLIG4a-R, 1824-2596 from *LIG4b* for pLIG4b-R and 3-498 from *XRCC4* for pXRCC4-R (coordinates in nucleotides from the ATG start codon of each gene). Control RNAi plasmids were p0ND7c (Garnier et al., 2004) and pICL7a (Gogendeau et al., 2008), which target *ND7* and *ICL7a* non essential genes, respectively. RNA interference was achieved as described (Galvani and Sperling, 2002), by feeding *Paramecium* cells with *Escherichia coli* HT115 bacteria transformed with each plasmid and induced for the production of double strand RNA corresponding to each RNAi insert.

Quantification of nuclear DNA content

During conjugation, 20 cells from each time-point were fixed and treated with RNase as described above and nuclear DNA was stained by VECTASHIELD with propidium iodide.

For DNA quantification, cells were flattened as described above and observed with a standard fluorescent microscope. To compensate for staining differences between slides, the fluorescence signal from new MACs was normalized relative to surrounding old MAC fragments (in which DNA has stopped to replicate but is not degraded under the experimental conditions used). DNA content in the new developing MAC (D) was calculated as follows:

$$D = \frac{(N-B)}{(O-B)} \times A$$

where N is the mean pixel intensity measured in the new MAC, B is background intensity, O is an average of fluorescence intensity in three old MAC fragments and A is the new macronucleus two-dimensional area. Since cells were well flattened, two-dimensional measurement was sufficient for this case. For each time-point, 8 to 15 cells were used for measurement.

References

Garnier, O., Serrano, V., Duharcourt, S., and Meyer, E. (**2004**). "RNA-mediated programming of developmental genome rearrangements in Paramecium tetraurelia." *Mol Cell Biol*, 24(17): 7370–9.

Gogendeau, D., Klotz, C., Arnaiz, O., Malinowska, A., Dadlez, M., de Loubresse, N. G., Ruiz, F., Koll, F., and Beisson, J. (**2008**). "Functional diversification of centrins and cell morphological complexity." *J Cell Sci*, 121(Pt 1): 65–74.

Timmons, L. and Fire, A. (**1998**). "Specific interference by ingested dsRNA." *Nature*, 395(6705): 854.

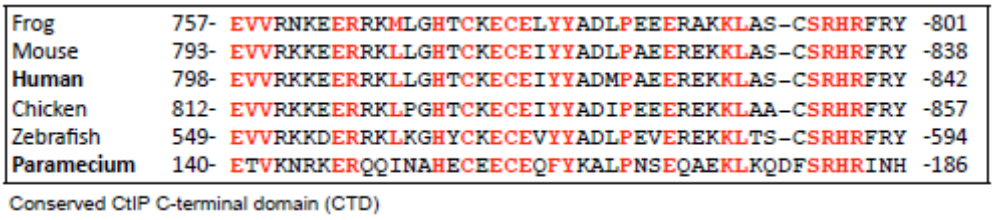


Figure 35 : Alignement des domaines C-terminaux de CtIP. Les résidus 100% conservés sont indiqués en rouge.

Résultats supplémentaires sur la maturation des extrémités.

Durant les réarrangements programmés du génome, les cassures double brin programmées sont introduites par la transposase domestiquée Pgm. Une réparation précise au nucléotide près est essentielle. On a vu que cette réparation est assurée par la voie NHEJ et implique une maturation contrôlée des extrémités cassées de l'ADN. D'une part l'élimination du dernier nucléotide en 5' au niveau de l'extrémité sortante. Et la polymérisation qui permettra de combler le « gap » induit par l'élimination de ce nucléotide.

Nous avons voulu savoir quel pourrait être la protéine responsable de cette activité de résection contrôlée.

Identification d'homologues de CtIP et de Mre11 dans le génome de Paramecium tetraurelia.

Deux homologues de CtIP ont été identifiés dans le génome de *Paramecium tetraurelia* par l'équipe de Alessandro Sartori (figure 35). Pour mener des analyses biochimiques, ils voulaient pouvoir profiter d'une protéine CtIP de petite taille, semblant réduite au domaine portant l'activité catalytique. Mais avant de pouvoir travailler sur cette protéine, il fallait apporter la preuve que CtIP était bien impliquée dans le mécanisme de recombinaison homologue. Un moyen facile de contrôler cela était de s'intéresser à la méiose de *Paramecium tetraurelia*. Une extinction de CtIP affecte-t-elle les processus sexuels ? De notre coté, nous nous demandions si CtIP ou Mre11 pouvaient être impliqués dans la résection des extrémités 5' de l'ADN, et plus particulièrement à l'élimination du nucléotide en 5' au niveau des extrémités cassées de l'ADN lors de l'excision des IES.

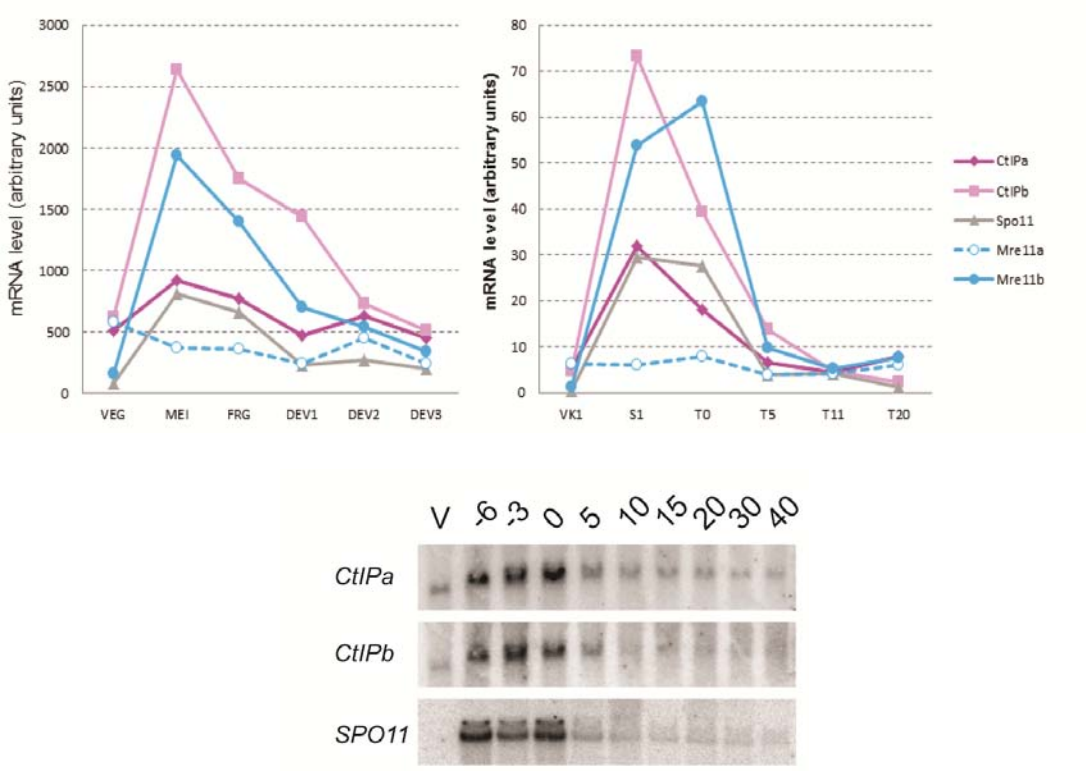


Figure 36 : Profil d'expression des gènes CtIP, MRE11 et SPO11. En haut à gauche, données de microarray. En haut à droite, données de RNA-seq. En bas, Northern Blot sur une cinétique d'autogamie.

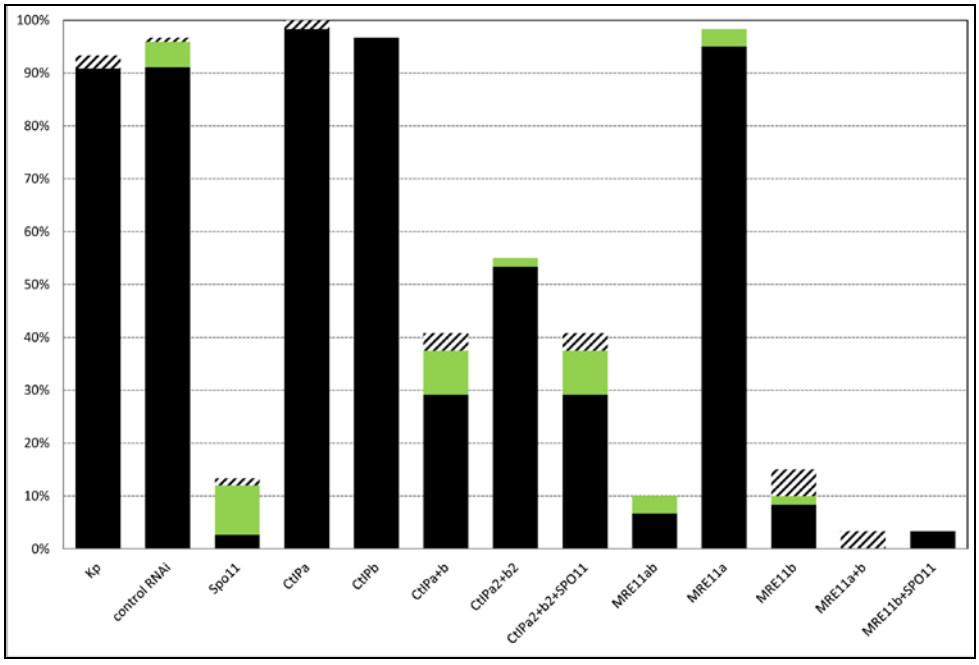


Figure 37 : Survie de la descendance lors de l'extinction de CtIP, MRE11 et SPO11. En noir, les vraies post autogames. En noir hachuré, les cellules « malades ». En vert, les cellules ayant régénéré l'ancien MAC.

Deux homologues de chacun de ces gènes ont été identifiés. Nous avons regardé le profil d'expression de ces gènes lors de l'autogamie (figure 36).

De façon attendue, les gènes *CtIPa*, *CtIPb*, *MRE11b* ont un pic de transcription dans les données de microarray, au stade méiose, exactement en même temps que *SPO11*, qui code pour la nucléase qui introduit spécifiquement des cassures double brin programmés pendant la méiose. *MRE11a*, en revanche, garde un niveau de transcription constitutivement bas. Le profil de transcription a été confirmé en ce qui concerne les gènes *CtIP* et Spo11 par Northern Blot et effectivement ces gènes ont bien un pic de transcription pendant les premiers temps de l'autogamie.

PtCtIP est nécessaire pour le passage de l'autogamie.

L'analyse du rôle de ces gènes a débuté par l'extinction systématique de tous ces gènes par ARN interférence, seuls, ou en combinaison (figure 37). Aucun effet n'a été observé lors de l'extinction des gènes *CtIP* individuellement mais l'extinction simultanée de *CtIPa* et *CtIPb* entrainait une mortalité de l'ordre de 45 à 60% dans la descendance post autogame. Cela nous indique que les deux gènes *CtIP* sont nécessaires et qu'ils ont probablement des fonctions redondantes chez *Paramecium* pendant l'autogamie.

A l'inverse l'extinction de *MRE11b* induit une très forte mortalité, de 90%, de la descendance post autogame, l'extinction de *MRE11a* n'a lui pas d'effet significatif par rapport à l'extinction contrôle d'un gène non essentiel.

Enfin l'extinction de *SPO11* induit également une forte mortalité, d'environ 85%, de la descendance post autogame.

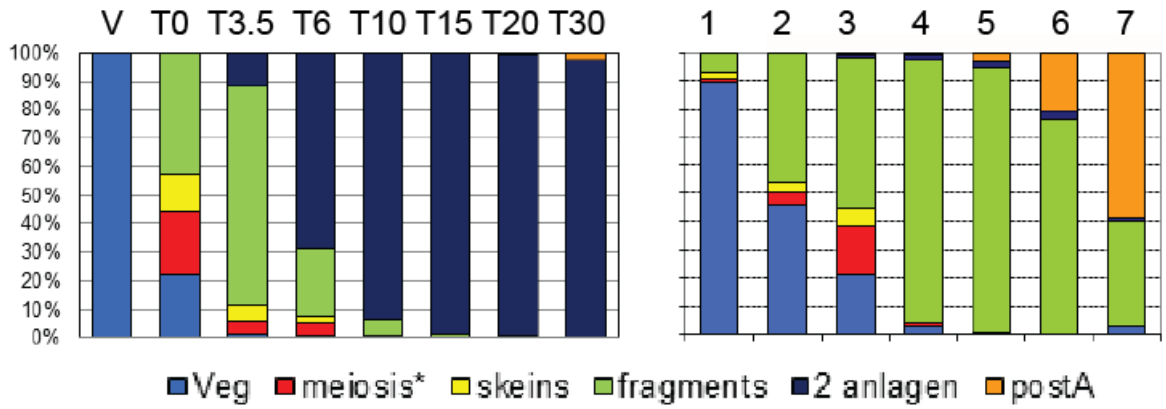


Figure 38 : Histogrammes des stades cellulaires pendant une cinétique d'autogamie contrôle (à gauche) ou lors d'une déplétion de CtIP (à droite).

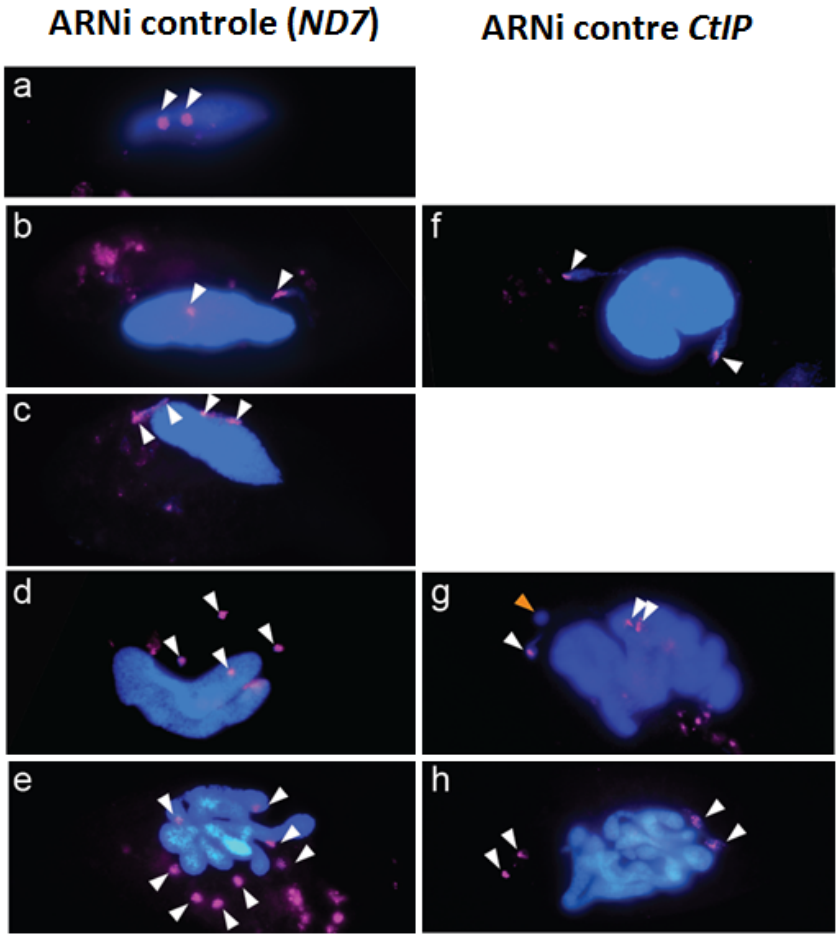


Figure 39 : Arrêt de la progression de la méiose dans les cellules déplétées de *CtIP*. Les mics (indiqués par des flèches) ont été marqués avec l'anticorps anti-CenH3, les noyaux au DAPI L'extinction de CtIP induit des problèmes à la méiose chez *Paramecium tetraurelia*.

Pendant l'autogamie de cellules déplétées simultanément de *CtIPa* et *CtIPb*, de *MRE11a* ou de *MRE11b* celles-ci ont été marquées au DAPI pour analyser le développement des nouveaux MACs dans la descendance (figure 39). De façon frappante, lorsque *CtIP* ou *MRE11b* sont éteints, une grande partie des cellules autogames présentent un stade « ancien MAC fragmenté » mais seulement très peu un stade « MAC en développement ». De même, on ne pouvait pas détecter les mics dans ces cellules. A des temps tardifs de l'autogamie apparait d'un seul coup dans la population de nombreuses cellules avec deux mics et un unique gros noyau entouré par de nombreux fragments de l'ancien MAC.

L'absence complète de cellules avec de nouveaux noyaux mics et deux MACs en développement lors de l'extinction de *CtIP* ou de *MRE11b* suggère un défaut précoce dans la formation du noyau zygotique. Chez *Tetrahymena thermophila*, une extinction de l'homologue de *CtIP* résulte dans le blocage de la méiose du micronoyau, du fait d'un défaut de réparation des cassures méiotiques introduites par Spo11, qui induisent des problèmes d'appariement et de disjonction des chromosomes homologues pendant la méiose.

Des anticorps anti CentH3, dirigés contre le variant centromérique de l'histone H3 de *Paramecium tetraurelia*, ont été utilisé pour marquer spécifiquement les chromosomes mic pendant la méiose de cellules déplétées de *CtIP*. Cela a permis de confirmer que la méiose est bloquée au stade « 4 mics » (correspondant à la première division de méiose) dans la grande majorité des cellules, alors que la fragmentation de l'ancien MAC progresse normalement.

La délétion du gène *SPO11* abolit le phénotype d'une extinction de *CtIP* ou de *MRE11b*

Chez la levure, un mutant Δsae2 n'est pas capable de former les tétrades pendant la méiose, mais la sporulation est restaurée dans un double mutant Δspo11 Δsae2, bien que les spores produites ne soient pas viables.

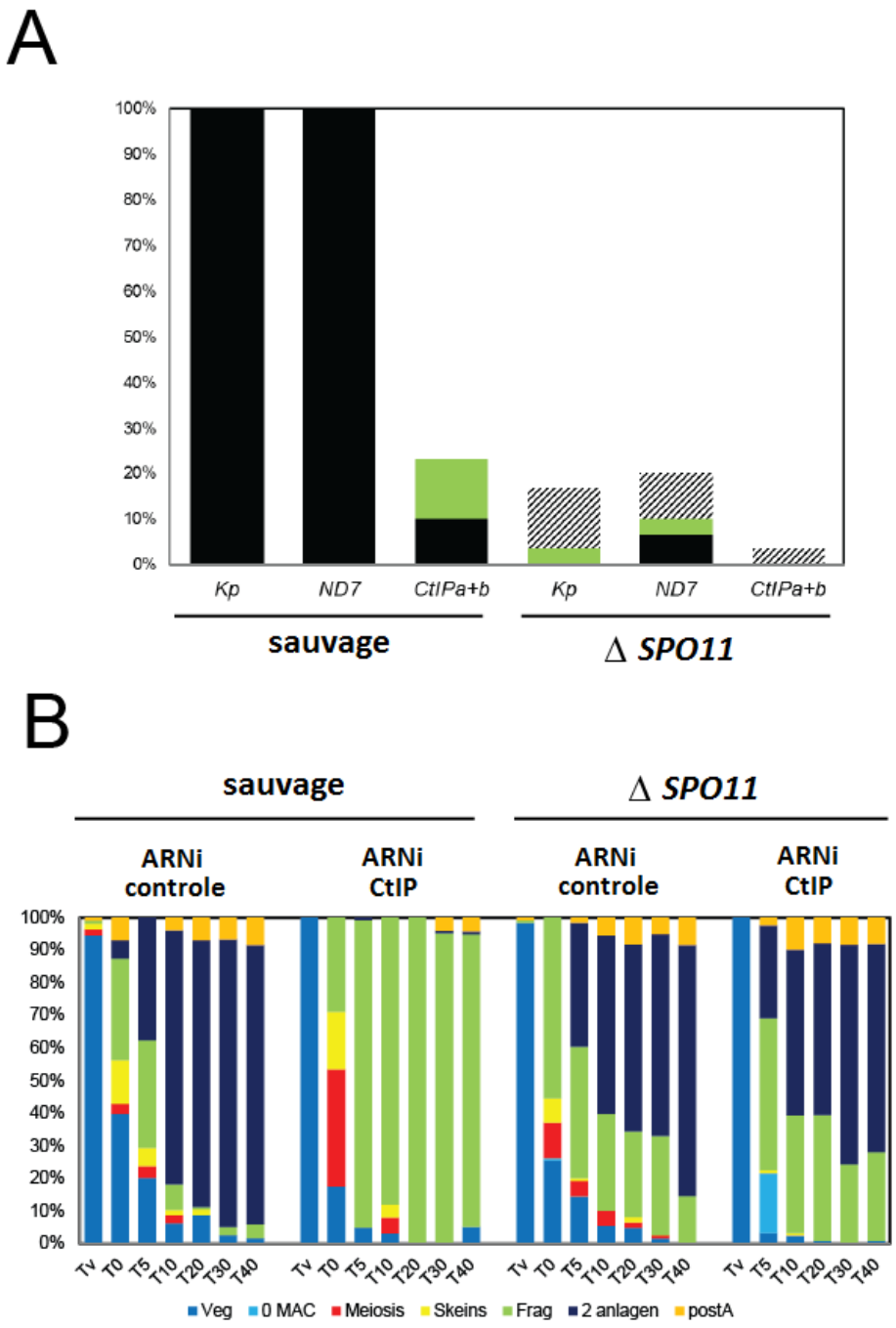


Figure 40 : Restauration du développement MAC dans les cellules ΔSPO11 lors de l'extinction de CtIP. A. Survie de la descendance. B. Suivi des stades cellulaires lors de cinétiques d'autogamie.

Pour aller plus loin dans l'analyse des homologues de *CtIP* de paramécie, nous avons induit au laboratoire la délétion somatique du gène *SPO11*. Après avoir vérifié que la délétion de *SPO11* induisait la mortalité de la descendance, probablement du fait de problème de ségrégation des chromosomes, nous avons fait passer l'autogamie à ces cellules déplétées d'un gène non essentiel ou de *CtIPa* et *b*. Dans ce dernier cas, la formation de nouveaux MAC est restaurée dans la souche Δspo11 + ARNi contre *CtIPa* et *CtIPb*. Le meme phénomène est observée dans les souches Δspo11 + ARNi contre *MRE11b* (figure 40). Cela indique que les homologues de paramécie de *CtIP* et *Mre11b* sont impliqués dans la réparation des cassures méiotiques induite par Spo11.

Lors de la réparation des cassures double brin méiotiques, CtIP participe avec l'endonucléase Mre11 à l'élimination des adduits ADN-Spo11 à l'extrémité 5'. Cette étape initie la résection de 5' vers 3', requise pour la réparation par recombinaison homologue.

Effet de CtIP sur l'excision des IES

L'effet de la déplétion de *CtIPa* et *CtIPb* sur la progression de la méiose pouvant être aboli par une délétion du gène *SPO11*, nous nous sommes intéressés à la résection du dernier nucléotide en 5' des extrémités de 4 bases sortantes produites par l'introduction des cassures double brin par Pgm (figure 41). J'ai réalisé des expériences classiques de LMPCR dans les cellules délétées dans le MAC du gène *SPO11*, avec ou sans déplétion de *CtIP*. Aucune différence significative dans l'introduction des cassures double brin programmée n'a été observée dans les différentes conditions. Cependant, pour l'IES 51A2591, situé dans le gène codant pour l'antigène de surface A, une électrophorèse haute résolution sur gel dénaturant a révélé la présence d'extrémités 5' sortantes non maturée (figure 42).

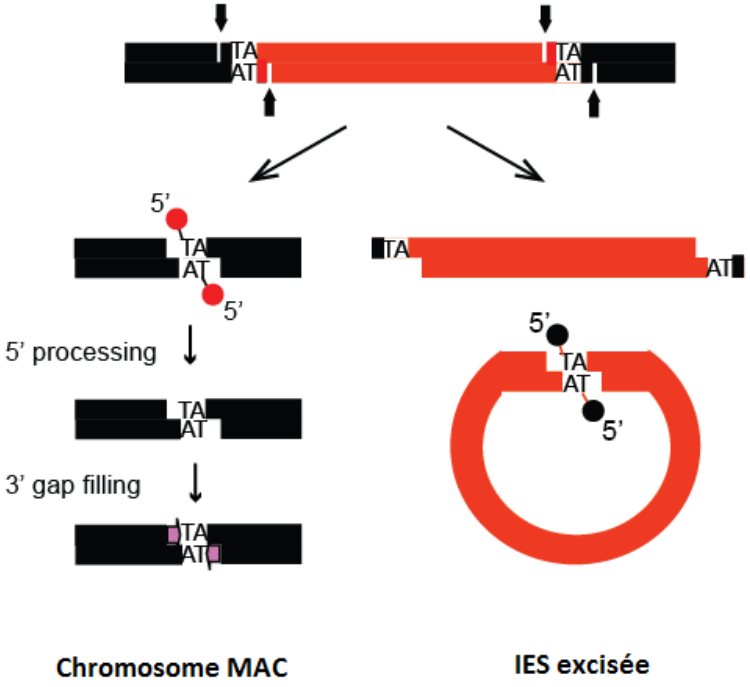


Figure 41 : Modèle d'excision des IES et maturation des extrémités ADN. Les IES apparaissent en rouge et l'ADN flanquant en noir.

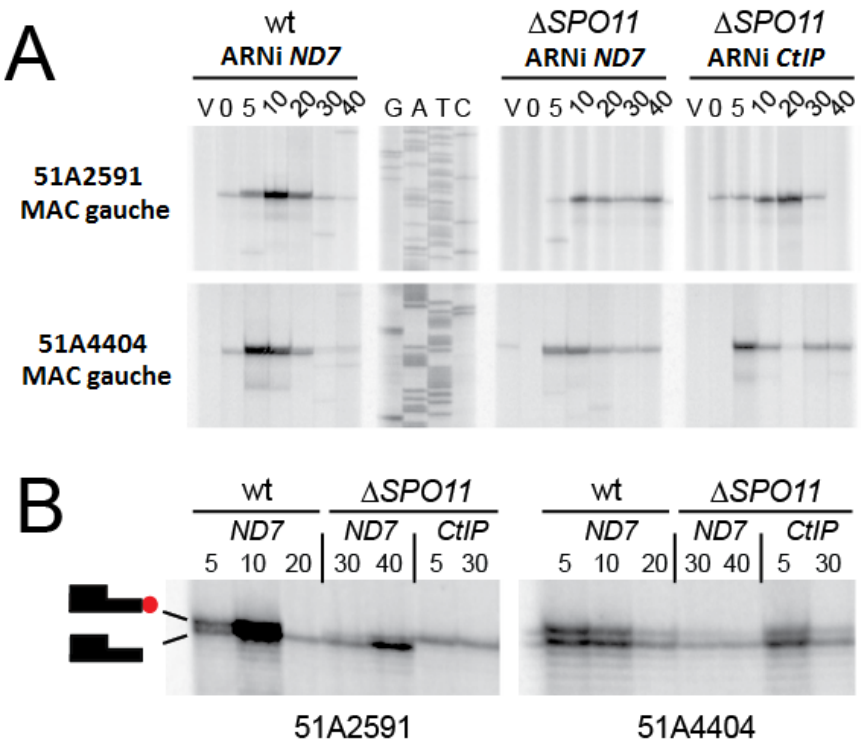


Figure 42 : Maturation des extrémités 5' au niveau des extrémités MAC flanquantes visualisée par LMPCR. A. Cinétiques. B. Les mêmes produits avec migration haute résolution.

Cette observation suggère que CtIP pourrait jouer un rôle dans l'élimination du nucléotide en 5' au niveau des extrémités sortantes de l'ADN MAC flanquant lors de l'excision des IES, avant réparation par la voie NHEJ. Cependant cette observation ne peut pas etre généralisée à toutes les IES. En effet, une maturation normale des extrémités a été observée pour l'IES 51A4404. Le même genre d'expérience sur d'autres IES seront nécessaires pour confirmer ces premiers résultats.

PAPIER KU + résultats supplémentaires

Role de Ku dans les réarrangements programmés.

Couplage entre introduction et réparation des CDBs

L’étude sur la Ligase IV et son partenaire Xrcc4 a permis de confirmer l’implication de la voie du NHEJ dans l’excision précise des IES. En effet ces protéines sont spécifiques de cette voie. Et lorsqu’elles sont absentes, les extrémités d’ADN non réparées s’accumulent mais ne semblent pas dégradés. Le rôle de Ku dans la voie canonique du NHEJ est de protéger les extrémités d’ADN double brin et de recruter les acteurs en aval de la voie. On imagine donc que Ku est aussi impliqué dans l’excision des IES, protège les cassures introduites par Pgm et recrute les facteurs de maturation puis le complexe Xrcc4 et LigIV pour l’étape finale de ligation.

Voie de réparation Ku indépendante, le alt-NHEJ.

Il a été décrit dans la littérature qu’il existait des voies de réparation indépendante de l’hétérodimère Ku. Elles sont caractérisées par des délétions de relative petite taille et l’usage de microhomologies pour diriger la réparation. Il en résulte une réparation imprécise.
L’hypothèse lors de mon arrivée au labo était de tester si des voies différentes étaient mobilisées pour l’excision des IES d’une part et pour l’élimination des séquences répétées d’autre part, et si la présence de Ku était le facteur déterminant entre réarrangements précis et imprécis chez *Paramecium tetraurelia*.

Pour mener cette étude j’ai donc entrepris l’extinction systématique des gènes Ku de la paramécie pour vérifier que l’hétérodimère était bien impliqué dans l’excision des IES et je devais mettre au points des outils me permettant d’étudier les réarrangements imprécis, bien moins étudiés et caractérisés que les mécanismes d’excision des IES.

J’imaginais qu’en l’absence des Ku, l’excision précise des IES seraient caractérisées par une imprécision au niveau des jonctions macronucléaires tandis que l’élimination imprécise des séquences répétées ne serait pas affectée.

Couplage entre introduction programmées des cassures et réparation par la voie NHEJ

Il est apparu au cours de la thèse que cette hypothèse de départ était fausse et que les réarrangements programmés du génome chez *Paramecium tetraurelia* mettent en jeu un couplage entre l’introduction programmée des cassures double brin par Pgm et la réparation de ces cassures par la voie NHEJ classique, Ku dépendante.

Ku-mediated coupling of DSB introduction and repair during programmed genome rearrangements

Antoine Marmignon [1], Julien Bischerour [1], Aude Silve [1], Clémentine Fojcik [1], Emeline Dubois [1], Vinciane Régnier [1,2], Olivier Arnaiz [1], Aurélie Kapusta [1,3], Sophie Malinsky [4,2], Mireille Bétermier [1]*

[1] CNRS UPR3404 Centre de Génétique Moléculaire, 1 avenue de la Terrasse, Gif-sur-Yvette F-91198 cedex, France; Université Paris-Sud, Département de Biologie, Orsay, F-91405, France

[2] Université Paris Diderot, Sorbonne Paris Cité, UFR Sciences du Vivant, F-75205 Paris, France

[3] Present address: University of Utah, Department of Human Genetics, 84112 Salt Lake City, UT, USA

[4] Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, F-75005 France; INSERM, U1024, Paris, F-75005 France; CNRS, UMR 8197, Paris, F-75005 France

* Corresponding author: mireille.betermier@cgm.cnrs-gif.fr

RUNNING TITLE: Ku couples DNA cleavage and DSB repair

SUMMARY

During somatic differentiation, physiological DNA double-strand breaks (DSB) can drive programmed genome rearrangements (PGR). The preservation of genome integrity, therefore, requires the mobilization of efficient DSB repair pathways. In the ciliate *Paramecium tetraurelia*, massive PGR involve the precise excision of 45,000 single-copy non-coding germline sequences (Internal Eliminated Sequences, IES) during the development of the somatic nucleus. A domesticated transposase, PiggyMac (Pgm), cleaves DNA at IES boundaries and IES excision sites are repaired through classical non-homologous end-joining (C-NHEJ). *P. tetraurelia* encodes two Ku70 and three Ku80 paralogs. We show here that a development-specific Ku heterodimer is essential for Pgm-dependent DNA cleavage *in vivo*. Pgm and Ku70/Ku80c both localize in the developing somatic nucleus and co-purify when overproduced in a heterologous system. Our study in *P. tetraurelia* provides the first evidence that Ku can be integrated in a DNA cleavage factory to ensure efficient coupling between DSB introduction and repair during PGR.

INTRODUCTION

DNA double strand breaks (DSBs) may arise accidentally in chromosomes, upon endogenous or exogenous stress. These lesions are among the most deleterious DNA damage: if left unrepaired, a single DSB may trigger cell death, while incorrect repair can give rise to chromosome rearrangements (reviewed in Chapman et al., 2012). Cells rely on two major pathways to repair DSBs: homologous recombination (HR) uses a homologous substrate to restore the sequence of the broken chromosome, while non-homologous end joining (NHEJ) proceeds through the ligation of free DNA ends. Even though they can be very toxic, programmed DSBs are obligatory intermediates in essential biological processes, such as meiosis or acquired immune response. During meiosis, the Spo11 endonuclease cleaves DNA and DSB repair is carried out by HR (reviewed in Longhese et al., 2009). In addition to favoring the exchange of parental alleles, HR ensures that homologous chromosomes are correctly paired before they are segregated during the first meiotic division. During lymphocyte differentiation, programmed genome rearrangements (PGR) mediated through V(D)J recombination generate the large diversity of immunoglobulin genes (reviewed in Schatz and Swanson, 2011). During V(D)J recombination, the domesticated transposase RAG1, associated with its partner RAG2, cleaves specific recombination sites. The resulting DSBs are repaired through classical NHEJ (C-NHEJ). A critical step in C-NHEJ is the binding of the Ku70/Ku80 heterodimer to broken DNA ends (reviewed in Lieber, 2010). Upon binding, Ku protects DNA ends from extensive resection and, together with its facultative partner DNA-PKcs, facilitates the synapsis of two broken ends. Following recruitment of DNA processing enzymes, the Ligase IV-Xrcc4 complex mediates the joining of DNA ends. An alternative end joining pathway, referred to as alt-NHEJ (or MMEJ, for microhomology-dependent end joining), has been reported (reviewed in McVey and Lee, 2008). This poorly characterized pathway is independent of Ku and, to some extent, of Ligase IV. Because of the absence of Ku, alt-NHEJ involves limited 5’ to 3’ resection of broken DNA ends and generates deletions at the repair sites, which often involve microhomologies. The initiation of alt-NHEJ differs from HR, because more extensive 5’ to 3’ resection takes place during HR to generate the long 3’ single strand that will invade a homologous DNA duplex (reviewed in Symington and Gautier, 2011).

Ciliates provide extraordinary models to study the molecular mechanisms involved in PGR (reviewed in Chalker and Yao, 2011). In these unicellular eukaryotes, two kinds of nuclei

coexist in the same cytoplasm. The highly polyploid somatic macronucleus (MAC) is essential for gene expression but it is destroyed at each sexual cycle, while the diploid micronucleus (MIC) undergoes meiosis and transmits the germline genome to the new MIC and MAC of the next generation. In *Paramecium*, massive PGR take place in the new developing MAC, while the genome is amplified from 2n to 800n (reviewed in Dubois et al., 2012). These PGR consist in the elimination of two types of germline-specific DNA. Regions of up to several kbp in length, often containing repeated sequences, are eliminated in a heterogeneous manner, leading to chromosome fragmentation or intra-chromosomal deletions. In addition, thousands of single-copy, short and non-coding Internal Eliminated Sequences (IES) are excised precisely. Because 47% of genes are interrupted by at least one IES in the germline genome (Arnaiz et al., 2012), the precise excision of IESs is essential for the assembly of functional genes in the new MAC and the survival of the sexual progeny. *Paramecium* IESs are invariably flanked by one TA dinucleotide on each side and little additional information can be found in their nucleotide sequence, which raises the question of how these sequences are recognized and targeted for excision. In fact, the excision of at least one third of *Paramecium* IESs is controlled maternally through a sequence homology-dependent mechanism (reviewed in Coyne et al., 2012). Indeed, a genome-wide comparison of the germline (from the old MIC) and rearranged (from the old MAC) versions of the genome is mediated by non-coding RNAs (Lepère et al., 2008; Lepère et al., 2009). This epigenetic control drives the trans-generational inheritance of rearrangement patterns, from the old to the new MAC.

In *Paramecium*, IES excision proceeds through a two-step “cut-and-close” mechanism. A domesticated *piggyBac* transposase, PiggyMac (Pgm), is essential to introduce the DNA cleavages that initiate the reaction, generating 4-bp staggered DSBs centered on the conserved TA at each IES boundary (Baudry et al., 2009). Following IES release, precise DSB repair is carried out through the C-NHEJ pathway and leaves a single TA at the IES excision site (Kapusta et al., 2011). Providing their length allows enough DNA flexibility, the excised linear IESs are circularized through the same DSB repair pathway, before they are actively degraded. During DSB repair, the flanking broken ends are thought to anneal through the pairing of the complementary TAs carried by their 4-base 5’ overhangs, and undergo limited 5’ and 3’ processing (Gratias and Bétermier, 2003). The final ligation step is mediated by the NHEJ-specific ligase complex, Ligase IV and its partner Xrcc4, both of which are essential

for PGR. When *LIG4* or *XRCC4* genes are knocked down, Pgm-dependent DSBs are introduced normally, but they accumulate in the developing new MAC, which severely compromises DNA amplification and impairs the recovery of a viable progeny (Kapusta et al., 2011).

Next-generation sequencing of the non-rearranged *Paramecium tetraurelia* genome has led to the identification of at least 45,000 IESs (Arnaiz et al., 2012). Therefore, a huge number of programmed DSBs are repaired precisely during MAC development. In the present study, we have addressed the question of how the C-NHEJ pathway is recruited to IES excision sites to carry out efficient and precise DSB repair. We have focused our analysis on the Ku heterodimer, which acts upstream in the C-NHEJ pathway, through its binding to DNA broken ends. Two *KU70* and three *KU80* genes were identified in the somatic genome (Kapusta et al., 2011). We report here that development-specific *KU* genes are essential for PGR. More surprisingly, we demonstrate that Ku is required for the introduction of programmed DSBs at IES boundaries, and provide evidence that Ku interacts functionally and physically with Pgm. We propose that the assembly of a complex including Pgm and Ku70/Ku80 is required to activate DNA cleavage during PGR in *P. tetraurelia*.

RESULTS

Development-specific *KU* genes in *P. tetraurelia*

Two genes encoding Ku70 homologs, *KU70a* and *KU70b*, were identified in the macronuclear genome assembly of *P. tetraurelia* {Figure 1A, \Kapusta, 2011 #482}. These two closely related copies arose from the more recent whole genome duplication (WGD) that took place during the evolution of this species (Aury et al., 2006), and are therefore referred to as ohnologs. We also identified three homologs of the human *KU86* gene (Figure 1A). *KU80a* and *KU80b* are ohnologs issued from the last WGD and the more distant *KU80c* has diverged after the intermediate WGD.

Sexual processes in *Paramecium* may occur in two different ways: during conjugation, following the mixing of two reactive partners with compatible mating types, or during a self-fertilization process called autogamy, in which MIC meiosis is induced in a single mating type upon starvation (reviewed in Bétermier, 2004). A microarray analysis of the *P. tetraurelia* transcriptome during autogamy has revealed that *KU70a* and *KU80c* are specifically induced during the development of the new MAC, when PGR take place (Arnaiz et al., 2010; Kapusta et al., 2011). Microarray data were confirmed through the analysis of high-throughput sequencing of polyadenylated RNAs extracted during an autogamy time-course (Figure 1B, Arnaiz et al. in prep.), northern blot hybridization with specific probes (Figures 1C and 3C, Malinsky et al. in prep.) and semi-quantitative RT-PCR (not shown). In a *PGM* knockdown (KD), the transcription of *KU70* and *KU80c* is switched on normally during autogamy, indicating that *KU* genes are not induced as a response to Pgm-induced DSBs, but more likely as part of a general transcription program during MAC development. Moreover, in contrast to control cells, the levels of *KU70* and *KU80c* mRNAs do not decrease at later time-points in a *PGM* KD, suggesting that the completion of PGR is a signal for transcriptional switch-off.

Ku70a and Ku80c localize in the developing new MACs

KU70a and *KU80c* expression is specifically induced during autogamy and reaches a peak when the new developing MACs start to be detected in the culture (T5 and T11 time-points in Figure 1B). Using N-terminal GFP fusions, we followed the cellular localization of Ku70a and Ku80c during autogamy. Transgenes expressing each fusion protein under the control of

their respective endogenous promoter were microinjected into the MAC of vegetative cells. The resulting transformants were grown and starved to induce autogamy. Consistent with the transcriptome analysis, GFP fluorescence was stronger in autogamous cells with respect to vegetative cells for each fusion transgene (Figure 2A). Both fusion proteins accumulated in the developing new MACs and concentrated into nuclear foci.

The detection of nuclear foci in the developing new MAC has been reported for a Pgm-GFP fusion protein in living cells, but the nature and biological significance of these foci has remained unclear (Baudry et al., 2009). We confirmed this localization in fixed and living cells, using a novel construct expressing a functional C-terminal Pgm-GFP fusion able to complement a somatic deletion of the *PGM* gene (Dubois et al., 2012 and unpublished data). Pgm foci were also observed when *RFP* was substituted for *GFP* in our *PGM* C-terminal fusion constructs (Figure S2). To compare the localization of Pgm and Ku foci, the *PGM-RFP* construct and the *GFP-KU80c* transgene were microinjected together into the MAC of vegetative cells. During autogamy of transformed cells, the Pgm and Ku80c fusion proteins colocalized in the same foci within the developing new MACs (Figure 2B), suggesting a possible involvement of Ku in Pgm-dependent PGR during MAC development.

Ku70 and Ku80c are required for the successful completion of autogamy

To test the implication of the different *KU* genes in MAC development, we knocked them down systematically by feeding wild-type *P. tetraurelia* cells on dsRNA-producing bacteria, (Galvani and Sperling, 2002). The very high percentage of identity between *KU70a* and *KU70b* made it impossible to design specific RNAi constructs for each individual gene. Therefore, we silenced them both together. In contrast, we designed gene-specific RNAi constructs for *KU80a, KU80b* and *KU80c.* Whenever possible, to make sure that the observed phenotypes would be attributable to the silencing of each targeted gene, we used RNAi constructs homologous to two different regions in each *KU80* gene (Figure S3). After three days of starvation in each silencing medium, individual autogamous cells were transferred to standard growth medium and allowed to resume vegetative growth. The survivors were grouped in three categories: (i) those that were able to undergo a new round of autogamy following a few divisions, most likely because they had regenerated their old MAC, were not considered as *bona fide* post-autogamous progeny (Kapusta et al., 2011); (ii) slow-growing survivors were counted as progeny with a defective new MAC; (iii) only the recovery of a

fully viable post-autogamous progeny indicated that the silenced cells had been able to form a functional new MAC (Figure 3A).

When RNAi was directed against *KU80a* or *KU80b*, either individually or together, the progeny exhibited good survival rates when compared to a control RNAi against the nonessential *ICL7* gene, or with respect to cells that underwent autogamy in standard medium (Figure 3A). In contrast, *KU80c* or *KU70* KDs yielded only 10 to 30% viable post-autogamous progeny in the following sexual generation. These data show that the developmentally-induced *KU80c* gene and one or both *KU70* genes are essential for full recovery of viable progeny.

The residual survival observed reproducibly in the *KU80c* and *KU70* KDs contrasted with the complete absence of survivors in the *XRCC4* KD (Figure 3B). This may reflect a different requirement for Ku and the Ligase IV/Xrcc4 complex during MAC development, as suggested by the apparently normal amplification of DNA in the developing new MACs of *KU80c* silenced cells, relative to cells silenced for *LIG4* or *XRCC4* (Figure 3B and Kapusta et al., 2011). The "faint DAPI-staining" phenotype of a *LIG4* KD is suppressed in a double *KU80c/LIG4* KD, confirming that Ku acts upstream of the Ligase IV during MAC development. The higher recovery of viable progeny in the *KU* KDs versus a *XRCC4* KD may also be attributable to the different gene expression profiles and to a lower efficiency of the feeding procedure to knock down genes that are induced at later stages. Indeed, *XRCC4* is induced during meiosis and can readily be silenced (Kapusta et al., 2011), while for *KU80c,* which is expressed later during autogamy, high amounts of full-length mRNAs are restored at T30 and T40, even though degraded transcripts are conspicuous at all time-points (Figure 3C).

KU70 and *KU80c* are essential for the programmed elimination of germline DNA

We have previously shown that the NHEJ-specific Ligase IV-Xrcc4 ligation complex carries out the joining of broken DNA ends at IES excision sites (Kapusta et al., 2011). Strikingly, in a *LIG4* KD, unrepaired broken ends remain very stable until late time-points during autogamy, suggesting that Ku may protect them against degradation. To gain further insight into the role of Ku during IES excision, we performed a molecular analysis of PGR in Ku-depleted cells. We expected that DSBs would be introduced normally at IES boundaries, but that the broken DNA ends would not be repaired correctly.

We first focused our analysis on a *KU80c* KD. Using Southern blot hybridization, we monitored the excision of one particular IES, IES 51G4404 from the surface antigen G^{51} gene (Duharcourt et al., 1995), during a large-scale autogamy time course of *P. tetraurelia* cells submitted to RNAi against *KU80c* (Figure 4A). In a control RNAi (left panel), a major band corresponded to the IES^- molecules from the old MAC and to the newly rearranged molecules that are amplified in the developing new MAC. A higher molecular weight species corresponding to the non-excised IES^+ form was detected transiently during MAC development and disappeared at late time-points, as a result of IES excision. In contrast, the IES^+ form was clearly amplified during autogamy in the *KU80c* KD (right panel). Using a related strategy, we tested chromosome fragmentation at one particular fragmentation site located downstream of the G^{51} gene (Figure 4B). We observed that non-rearranged molecules are amplified during MAC development in the *KU80c* KD, which indicates that Ku is also required for the elimination of the germline DNA that is associated with chromosome fragmentation.

The amplification of non-rearranged DNA indicates that PGR do not proceed normally in a *KU80c* KD. However, because of the presence of old MAC DNA in our samples, the above experiment could not detect whether residual rearrangements had taken place in the new MAC. We could circumvent this problem, because the strain used in this experiment (51ΔA) harbors a wild-type germline genome, but carries a somatic deletion of the nonessential surface antigen gene A^{51}. During autogamy, all IESs are excised normally from the A^{51} gene before the whole locus is deleted from the new somatic MAC (Gratias et al., 2008). Therefore, in the 51ΔA variant, all the IES^- molecules detected at this locus during autogamy may be attributed to *de novo* IES excision in the new MAC. Using PCR primers hybridizing in the flanking sequences, excised (IES^-) and non-excised (IES^+) molecules were readily detected in a control RNAi experiment, for IESs belonging to different classes (Figure 4C, left panel): short (51A1835: 28 bp), intermediate (51A4404: 77 bp) or long (51A2591: 370 bp), maternally (51A2591) or non-maternally controlled (51A1835 and 51A4404). This stands in sharp contrast to the complete absence of *de novo* IES excision junctions in the *KU80c* KD (right panel). Likewise, using IES-specific internal divergent primers, we followed the appearance of excised IES circular molecules during the autogamy of control and *KU80c*-silenced cells, but no circle junctions were detected in the *KU80c* KD (Figure 4D).

To extend our analysis to the other *KU* genes, we submitted small-scale cultures of 51ΔA cells to RNAi against *KU70* or individual *KU80* genes, and tested their ability to complete IES excision during autogamy, using the above PCR assay for IESs within the A^{51} gene. When *KU80a* and *KU80b* were knocked down, together or separately, rearranged molecules appeared at day 2 of starvation for all IESs, with the same timing as in a control RNAi (Figure S4). In contrast, the appearance of *de novo* IES excision junctions was strongly inhibited in a *KU80c* or a *KU70* KD (Figure S4). Taken together, our molecular data indicate that the development-specific Ku70 and Ku80c proteins are essential for the recovery of both precise chromosomal junctions and excised IES circles during PGR. Moreover, in *KU70* and *KU80c* KDs, the non-rearranged version of the genome is amplified in the developing MAC, a strikingly different phenotype from that of *LIG4* or *XRCC4* KDs (Kapusta et al., 2011), but quite similar to a *PGM* KD (Baudry et al., 2009).

No DNA breaks are detected in cells depleted for Ku80c

The absence of precise *de novo* IES excision junctions in a *KU70* or *KU80c* KD could either reflect a problem in DSB repair, as established for a *LIG4* KD, or an inhibition of DNA cleavage, as demonstrated for a *PGM* KD. Because Ku is essential for C-NHEJ in all organisms, a defect in end-joining was expected in *KU* KDs in *P. tetraurelia*. However, the observation that non-excised IESs were amplified in the new MAC of cells depleted for Ku70/Ku80c raised the issue of whether DSBs were actually introduced at IES boundaries. We therefore used sensitive ligation-mediated PCR (LMPCR) to search for DSBs at IES boundaries in a *KU80c* KD. As previously published (Gratias and Bétermier, 2003), for those IESs that were tested, free broken DNA ends at IES boundaries were detected at early time-points, during the autogamy of cells submitted to a control RNAi (Figure 5A). DSBs disappeared later on during MAC development, indicative of their efficient repair. In a *KU80c* KD, we could detect no DSBs, neither on MAC ends nor on IES ends (Figure 5A).

Ku is known to protect broken DNA ends against degradation in other organisms (reviewed in Chapman et al., 2012), and a depletion in Ku proteins may reveal alternative DSB repair pathways, such as alt-NHEJ or HR, by allowing 5’ to 3’ DNA end resection. Should resection occur in *Paramecium KU* KDs, the resected 5’ DNA ends would not be appropriate substrates for LMPCR, because this technique only detects DSBs with a specific geometry (4- or 3-base 5’ overhangs). We first investigated whether alt-NHEJ might rescue *KU* KDs, by repeating

the PCR assays shown in Figure 4C with more distant primers hybridizing 1 kb away from the IES excision site. This would have allowed us to detect alternative repair junctions with small deletions attributable to alt-NHEJ. Even under these conditions, however, no heterogeneous excision junctions could be detected (not shown). Second, we reasoned that HR could use non-rearranged DNA molecules as substrates to repair DSBs at IES excision sites, which would account for the amplification of IES^+ molecules in the new MAC (Figure 4A). During HR-mediated DSB repair, the free 3'OH ends resulting from Pgm-dependent DNA cleavage should not be degraded and should be detectable through polynucleotidyl terminal transferase (TdT) tailing (Gratias and Bétermier, 2003). In a control RNAi, free 3'OH ends were detected transiently at IES boundaries (Figure 5B), with the same timing as the DSBs observed using LMPCR (Figure 5A). However, no free 3'OH ends were observed at the expected position in a *KU80c* KD (Figure 5B). Taken together, our data do not support the hypothesis that, in cells depleted for Ku, putative DSBs are repaired at IES excision sites through an alternative pathway involving 5' to 3' resection. They rather suggest that Ku is required before DNA cleavage, and that it may interact with Pgm before the introduction of DSBs at IES boundaries.

Functional and physical interaction between Ku and Pgm

The requirement for Ku before DNA cleavage could reflect an interaction between Ku and Pgm during the initiation of PGR. This hypothesis would be consistent with the observation that GFP-Ku80c and Pgm-RFP fusions colocalize in the developing new MACs (Figure 2B). To test whether a functional interaction exists between the two proteins, cells expressing a Pgm-GFP fusion were submitted to RNAi against *KU80c*, and the localization of the fusion protein was monitored during autogamy. In a control KD, the fusion protein appeared in the developing new MAC at day 2 of starvation, and formed small foci until day 3 (Figure 6A, top). The Pgm-GFP foci disappeared at day 4, which corresponds to full completion of PGR (see Figure S4). In the *KU80c* KD, the Pgm-GFP fusion was still produced and localized to the developing new MAC (Figure 6A, bottom), indicating that Ku is neither essential for the induction of *PGM* expression (Figure 6B), nor for the import of Pgm into the new MAC. However, a *KU80c* KD resulted in a dramatic increase in the nuclear concentration of the Pgm-GFP fusion (Figure 6A, bottom). At the RNA level, the *PGM* mRNA remained continuously produced throughout autogamy instead of being switched off at the latest time-points (Figure 6B). This change in transcription pattern parallels our previous observation that

KU70 and *KU80c* mRNA levels remain very high in a *PGM* KD (Figure 1C), and further supports our suggestion that the completion of PGR switches off the transcription of at least a subset of autogamy-specific genes. With regard to the subnuclear localization of Pgm, we also observed a striking difference in the *KU80c* KD with respect to the control. Indeed, at day 3, Pgm-GFP accumulated in large nuclear bodies, which were clearly detectable under Nomarski contrast and coincided with DAPI-free regions (Figure 6A). A similar aberrant localization has been observed for Pgm-GFP mutant proteins that are unable to restore IES excision in a *PGM* KD and, therefore, are deficient for PGR (E. Dubois, unpublished results). Likewise, in a *KU80c* KD, the appearance of large Pgm-GFP nuclear bodies may reveal a defect in DNA cleavage, which suggests that a functional interaction exists between Pgm and Ku during PGR. At day 4, the GFP fluorescence was still high, but the large nuclear bodies were not detected anymore. The apparently normal subnuclear organization of Pgm-GFP at day 4 coincides with the detection of low levels of IES excision (Figure S4), which, as stated above, might be explained by the restoration of functional amounts of *KU80c* mRNA at very late time-points (Figure 3C).

The functional interaction described above could rely on the formation of a protein complex containing Ku and Pgm. To investigate this hypothesis, HA-tagged versions of Ku70a or Ku80c were produced in a heterologous insect cell system, together with Pgm fused to the maltose-binding protein (MBP) at its N-terminal end (Figure 6C). For each condition, the MBP-Pgm protein was precipitated from soluble cell extracts using amylose magnetic beads, and we monitored the co-precipitation of the Ku subunits on western blots. We observed that both Ku70 and Ku80c were co-purified with MBP-Pgm, either individually or when the two subunits were co-expressed in the same cells. Control experiments using the MBP tag alone, or no MBP at all, confirmed that the enrichment in either Ku subunit was specific for the presence of Pgm in the extracts (Figure S5A). Reciprocally, we could co-immunoprecipitate Pgm with HA-tagged Ku proteins, using magnetic beads coated with anti-HA antibodies (not shown). The association of Pgm and Ku resisted to DNase treatment (Figure S5B), suggesting that the formation of a Pgm/Ku complex may be independent of the presence of DNA. Taken together, our data indicate that Pgm and Ku assemble in a higher-order complex in soluble cell extracts.

DISCUSSION

A development-specific Ku heterodimer is required for Pgm-dependent IES end cleavage

Gene duplication has been proposed to be a driver of genome evolution, allowing the sub-functionalization of duplicated genes and sometimes leading to the emergence of novel cellular functions (reviewed in Taylor and Raes, 2004). In *P. tetraurelia*, the presence of two *KU70* and three *KU80* genes has been the result of successive WGDs (Aury et al., 2006). The *KU80* family provides a nice example of sub-functionalization, with *KU80c* having acquired a characteristic induction pattern during MAC development, which correlates with its specific function during PGR. Future studies should address the question of whether the sub-functionalization of *KU80c* has resulted only from its overexpression or from the specialization of the protein. The function(s) of the *KU70* genes could not be investigated separately through RNAi. However, using RNA deep sequencing, we confirmed that the two recently duplicated *KU70* genes exhibit distinct transcription patterns, with *KU70a* being overexpressed during MAC development. We propose that, similar to *KU80c*, *KU70a* might have been undergoing specialization to carry out an essential function in PGR.

In *KU70* or *KU80c* KDs, both IES excision and chromosome fragmentation are inhibited. However, we confirmed that *PGM* expression, which is essential for the two types of genome rearrangements (Baudry et al., 2009), is induced normally in a *KU* KD, and that Pgm still localizes to the developing new MACs. However, a LMPCR molecular study failed to detect DSBs with the expected geometry at IES boundaries. Consistent with a defective C-NHEJ pathway, no *de novo* precise IES excision junctions were detected in *KU* KDs. Neither did we detect any imprecise junction that may have resulted from alt-NHEJ. Instead, we found that the non-rearranged version of the genome is amplified in the new MAC. Because IES excision starts after 3 to 4 rounds of genome amplification in the new MAC, we considered the possibility that HR substitutes for end-joining during the repair of IES excision sites. HR-mediated DSB repair would restore IES^+ chromosomes, providing a yet non-rearranged sister chromatid were used as a template. However, using a sensitive TdT-tailing assay, we obtained no evidence that HR intermediates are formed, based on the absence of detectable free 3’ ends at IES boundaries, which would have been missed by LMPCR due to 5’ to 3’ resection. Our data, therefore, point to the participation of the development-specific Ku70-Ku80c

heterodimer in Pgm-dependent DNA cleavage, upstream of its likely function in DSB repair as a core C-NHEJ protein.

A developmental transcription program in *Paramecium*

Previous analysis of the *P. tetraurelia* transcriptome has revealed that *KU80c* and *PGM* belong to the same "intermediate" cluster of genes that are induced during MAC development and repressed at later time-points during autogamy (Arnaiz et al., 2010). The transcription of *PGM* is induced normally in a *KU80c* KD. However, while *PGM* expression decreases at late time-points in a control, it is continuously turned on in a *KU80c* KD, while PGR are strongly inhibited. Reciprocally, in a *PGM* KD, where no PGR take place, the transcription of *KU80c* is induced normally but is not switched off. Similarly, we observed that in a *DNA-PKcs* KD, where PGR are also impaired, the transcription of *KU* and *PGM* genes fails to decrease at late stages of MAC development (Malinsky et al., in prep.). These observations suggest the existence of a feedback regulatory loop that would depend upon the completion of DNA rearrangements. For instance, a transcriptional activator specific for the intermediate gene cluster may be expressed from an IES-containing gene in the developing MAC, and switched off as soon as IES excision has been completed. A regulatory mechanism relying on the retention of an IES overlapping a gene promoter has been shown to control the expression of a mating-type gene in *P. tetraurelia* (Singh et al., submitted). In the ciliate *Euplotes crassus*, PGR also regulate the expression of a development-specific telomerase: the gene is localized in the germline-restricted part of the genome and is, therefore, switched off naturally once it is eliminated (Karamysheva et al., 2003). The existence of such regulatory loops provides a nice illustration of how ciliates may have taken advantage of PGR to fine-tune gene expression during their sexual cycle.

Coupling between DNA cleavage and DSB repair in ciliates

In another ciliate, *Tetrahymena thermophila*, IES excision is mediated by Tpb2p, a Pgm homolog that is responsible for DSB introduction at IES boundaries (Cheng et al., 2010). *T. thermophila* harbors one *KU80* and two *KU70* genes (Lin et al., 2012). In a *TKU80*Δ strain, Tpb2p-dependent DNA cleavage was reported to take place, but DSBs are not repaired, which eventually leads to an arrest in MAC development and to DNA loss in the new MAC. These phenotypes, which are quite similar to those described for a *LIG4* or *XRCC4* KD in *P. tetraurelia* (Kapusta et al., 2011), are fully consistent with a classical scenario, in which C-

NHEJ proteins are recruited to broken DNA ends after the introduction of DSBs at IES boundaries. The co-localization of Tpb2p and Tku80p in large heterochromatin bodies may reflect the assembly of nuclear factories that would facilitate the channeling of broken DNA ends towards C-NHEJ (Lin et al., 2012).

Noteworthy, IES excision is rather imprecise in *T. thermophila*, and heterogeneity at IES excision junctions could be attributed to variability in the choice of Tpb2p cleavage sites (Saveliev and Cox, 1995), or to the participation of different mechanisms in the formation of IES excision junctions, such as C-NHEJ if both IES boundaries are cut (Lin et al., 2012) or trans-esterification if a DSB is introduced at a single boundary (Saveliev and Cox, 1996, 2001). Accordingly, *T. thermophila* appears to have avoided IES insertion into coding regions (Fass et al., 2011), where imprecise excision might have deleterious effects. Furthermore, the presence of Ku is not a prerequisite for DNA cleavage in this ciliate. The situation is quite different in *P. tetraurelia*, in which a huge number of genes are interrupted by at least one IES (Arnaiz et al., 2012). In this ciliate, the pressure to assemble functional open reading frames in the somatic genome has driven the emergence of a highly efficient and precise IES excision mechanism. Several factors may have contributed to this precision: (i) the establishment of a crosstalk between IES ends before DNA cleavage, within a transpososome-like complex (Arnaiz et al., 2012; Gratias et al., 2008), (ii) the precise positioning of Pgm-dependent cleavages on each flanking TA (Gratias and Bétermier, 2003) and (iii) a tight coupling of DNA cleavage and C-NHEJ-mediated DSB repair through the incorporation of Ku into the DNA cleavage complex containing Pgm (Figure 7).

Here, we show that DNA cleavage and DSB repair are intertwined in *P. tetraurelia*. We propose that, in *Paramecium*, the development-specific Ku80c/Ku70a heterodimer is an essential partner of Pgm and that it activates DNA cleavage at IES boundaries. The incorporation of C-NHEJ proteins in the cleavage complex would allow the cell to face the challenge of repairing efficiently and precisely tens of thousands of DSBs in a short time during MAC development. We cannot exclude, at this stage, that the DNA-PKcs is also a component of the complex (Figure 7). Indeed, a *DNA-PKcs* KD impairs the excision of at least a subset of IESs and also likely reduces the efficiency of DNA cleavage at the boundaries of sensitive IESs (Malinsky et al., in prep.). In contrast, Ligase IV and Xrcc4 are clearly not required for DNA cleavage (Kapusta et al., 2011) and are probably not part of the cleavage complex.

Integration of DNA cleavage and repair in recombination factories during PGR

Domesticated transposases from cut-and-paste DNA transposons might have been recruited to perform PGR in different systems, not only because of their DNA cleavage activities, but perhaps also because of the peculiar features of cut-and-paste transposition (reviewed in Feschotte and Pritham, 2007). Indeed, when transposons integrate into a novel target site, they duplicate a short sequence, the TSD (target site duplication), on each side of the integrated element. During the next round of transposition, the transposase cuts the DNA, excises the transposon and leaves a DSB at the donor site, which is repaired through cellular pathways. The study of cut-and-paste transposons has revealed that C-NHEJ accounts for the characteristic footprints that are generated during DSB repair at transposon donor sites, with two copies of the TSD flanking a few bp from the transposon (Beall and Rio, 1996; Izsvak et al., 2004; Robert and Bessereau, 2007; Yant and Kay, 2003). Interestingly, several cut-and-paste transposons/transposases interact with Ku. For instance, Ku70 binds the ends of the *P* element from *Drosophila melanogaster* and stimulates DSB repair at the donor sites (Beall and Rio, 1996). *In vitro*, the transposase of *Sleeping Beauty*, a reconstructed *Tc/mariner* transposon, forms a complex with Ku70, and efficient transposition of *Sleeping Beauty* in a cellular system depends on the presence of DSB repair proteins (Izsvak et al., 2004). However, it is not clear *in vivo* whether the Ku/transposase interaction activates DNA cleavage at transposon ends or whether it simply facilitates the recruitment of the C-NHEJ pathway after excision.

With regard to PGR, the formation of a complex involving a nuclease and DSB repair factors has been hypothesized during V(D)J recombination (Schatz and Ji, 2011). *In vitro,* Ku interacts with the domesticated transposase RAG1 (Raval et al., 2008), but no evidence has been provided that Ku is required *in vivo* for the introduction of RAG1-dependent programmed DSBs. Similar to *Tetrahymena* (Lin et al., 2012), the formation of nuclear recombination centers, or recombination factories, may orchestrate V(D)J recombination through bringing together recombination sites marked by chromatin modifications, the endonuclease RAG1/RAG2 and DSB repair factors, but the presence of C-NHEJ proteins may not be required for DNA cleavage itself. The interplay between DNA cleavage and DSB repair has been pushed one step forward in *Paramecium*. Indeed, the present study of IES excision in *P. tetraurelia* provides the first evidence that Ku is absolutely required *in vivo* to introduce programmed DSBs during PGR. The demonstration that transposase-mediated

DNA cleavage and C-NHEJ mediated repair are coupled during PGR in *P. tetraurelia* parallels the observation that, during meiosis in *S. cerevisiae*, the Mre11p HR protein is required for DNA cleavage by the topoisomerase-like Spo11 endonuclease (reviewed in Borde, 2007). The two systems support the notion that the presence of DSB repair factors in recombination factories may be a prerequisite for DNA cleavage during programmed genome rearrangements.

EXPERIMENTAL PROCEDURES

P. tetraurelia strains and growth conditions

For autogamy time-course experiments, we used *P. tetraurelia* strain 51 new (hereafter called 51) and its 51ΔA variant carrying a heritable deletion of the *A* gene in its MAC but harboring a wild-type MIC (Gratias et al., 2008). To facilitate the screening of transformants in microinjection experiments, we introduced the *nd7-1* mutation (Skouri and Cohen, 1997) into strain 51 by conjugation, or used somatic variants carrying a MAC deletion of the *ND7* gene (Garnier et al., 2004). Cells were grown at 27°C in a wheat grass infusion (WGP; Pines International Inc.) inoculated with *Klebsiella pneumoniae*. Autogamy was carried out through starvation as described in (Bétermier et al., 2000). Total RNA and genomic DNA were extracted from ~400,000 cells for each time-point and quantified as described in (Baudry et al., 2009).

RNA interference by feeding

RNAi plasmids.

All RNAi plasmids are derivatives of vector L4440 and carry a target gene fragment between two convergent T7 promoters. The risk of cross-silencing was minimized by using the "RNAi-off-target" tool of ParameciumDB (Arnaiz and Sperling, 2011). A detailed description of all RNAi constructs can be found in the Supplementary Information.

RNAi during autogamy.

RNAi during autogamy was performed on strains 51 or 51ΔA as described in (Baudry et al., 2009). Survival of the progeny was tested at day 4 of starvation by transferring 30 individual autogamous cells to standard *K. pneumoniae* medium.

Molecular procedures

Oligonucleotides were purchased from Sigma-Aldrich or Eurofins MWG Operon (Table S1). PCR amplifications were performed in a final volume of 25 µL, with 10 pmol of each primer, 5 nmol of each dNTP and 1 U of DyNAzyme II DNA polymerase (Finnzymes) or DreamTaq (Thermo Scientific), using an Eppendorf Mastercycler personal thermocycler. PCR products were analyzed on 3% NuSieve GTG agarose gels (BioWhittaker Molecular Applications). LMPCR detection of double-strand breaks was performed as described in (Gratias et al., 2008). Poly(C) tailing of free 3′ ends using terminal transferase was performed as described in

(Gratias and Bétermier, 2003), using 500 ng of input total genomic DNA. All DNA sequencing was performed by GATC Biotech.
Northern and Southern blot hybridization with ^{32}P-labeled probes was carried out as described in (Baudry et al., 2009).

RNA-seq gene expression level

Strand-specific RNA-seq libraries were prepared from 50 ng polyA+ RNA following the directional mRNA-seq library preparation protocol provided by Illumina: RNAs were fragmented using fragmentation buffer, purified and treated with phosphatase and kinase prior to sequential ligation with different RNA adapters to the 3' and 5' ends. The ligated RNA fragments were reverse transcribed, followed by PCR amplification. The library was sequenced using an Illumina Genome Analyzer IIx to generate 75 nt paired-end reads. Reads were mapped with TopHat2 (Kim et al., 2013) (read-mismatches 1; min-intron-length 15; max-intron-length 100) on the *P. tetraurelia* strain 51 reference MAC genome (Arnaiz et al., 2012). Alignments were indexed using Samtools (Li et al., 2009). A custom perl script using the Bio::DB::Sam module was written to count the number of unique reads and fragments for each gene model. The counts were normalized to account for gene size and the number of mapped reads.

Paramecium transformation and localization of fluorescent fusion proteins

Plasmids carrying *GFP* or *RFP* fusion transgenes are described in the Supplementary Information. All plasmids were linearized by appropriate restriction enzymes and microinjected into the MAC of vegetative 51 *nd7-1* or 51ΔND7 cells, as described in (Baudry et al., 2009). A complementing plasmid carrying a functional *ND7* gene (Skouri and Cohen, 1997) was coinjected with the fusion transgenes. No lethality was observed in the post-autogamous progeny of transformed cells (data not shown), indicating that the GFP-Ku80c and GFP-Ku70a fusions did not interfere with the normal progression of autogamy. Cells were permeabilized for 4 min in PHEM (60 mM Pipes, 25 mM Hepes, 10 mM EGTA, 2 mM $MgCl_2$ pH 6,9) + 1% Triton, then fixed for 10 min in PHEM + 2% paraformaldehyde. All observations were performed using a Zeiss Axioplan 2 Imaging epifluorescence microscope. Developing MACs were identified using Nomarski differential interference contrast (DIC) combined with 4',6-diamidino-2-phenylindole (DAPI) staining.

Protein expression in insect cells and co-precipitation assays

Plasmids and vectors

DNA sequences coding for Pgm, Ku70a and Ku80c were obtained by gene synthesis (DNA 2.0 or Eurofins MWG/Operon). The synthetic *PGM* DNA sequence was first cloned into the pMAL-c2x vector (New England Biolabs). The *MBP-PGM* fusion, the *MBP*, the *KU70a* and the *KU80c* sequences were further cloned into the pVL-1392 vector (BD Biosciences). A HA tag was introduced at the C-terminus of Ku70a and the N-terminus of Ku80c. All plasmid sequences are available upon request. In a second step, plasmids pVL1392-MBP-Pgm, pVL1392-MBP, pVL1392-Ku70a-HA and pVL1392-HA-Ku80c were co-transfected individually into High Five™ cells with the BD BaculoGold™ Linearized Baculovirus DNA (BD Biosciences). Recombinant baculoviruses were further used for infection and protein expression in High Five™ cells (see details in the Supplemental Information). In this system, each gene is expressed under the control of a late viral gene.

Co-precipitation assays

All experimental details for the preparation of protein extracts from High Five™ cells infected with *MBP*, *MBP-PGM*, *KU70a-HA* and/or *HA-KU80c* recombinant baculoviruses can be found in the Supplemental Information. Soluble extracts were incubated for 2 hrs with 100 µg of amylose magnetic beads (New England Biolabs). Beads were washed 3 times with 1 ml of lysis buffer, then re-suspended in Laemmli buffer before electrophoresis in SDS-polyacrylamide gels. Aliquots were saved for input control. The co-precipitation of Ku proteins was detected on western blots, using anti-HA primary antibodies (monoclonal HA-7 from Sigma Aldrich) and anti-mouse HRP-coupled secondary antibodies (Promega). The MBP and MBP-Pgm proteins were detected with anti-MBP-HRP coupled antibodies (New England Biolabs).

FIGURE LEGENDS

Figure 1: *KU70* and *KU80* genes in *P. tetraurelia*

A. *KU70* and *KU80* genes in *P. tetraurelia*. Top: the black triangles represent each whole-genome duplication. Bottom: the % of identity between genes (nt) and proteins (aa) are indicated along the arrows connecting two genes.

B. Transcription profiles of *KU70* (left) and *KU80* (right) genes, as determined by high-throughput RNA Seq of samples extracted during an autogamy time-course of strain 51. Vk: vegetative cells; S1: starved or meiotic cells with intact parental MAC; T0: 50% of cells with fragmented MAC; the following time-points refer to hours after T0 (Nowak et al., 2011). On the vertical axis, FPKM represents the number of fragments per gene kb per million of reads uniquely mapped on the genome.

C. Detection of *KU80c* mRNA during autogamy, through northern blot hybridization. Control time course experiment: RNAi against the nonessential *ND7* gene (Skouri and Cohen, 1997), which encodes an exocytosis protein. *PGM*: RNAi against *PGM*. The *KU80c* probe is shown in Figure 3A. V: vegetative cells; T0 (not shown) corresponds to 50% of cells with fragmented MAC; the time-points refer to hours after T0 (Figure S1A for details).

Figure 2: Localization of Pgm and Ku GFP fusions during autogamy

A. Localization of GFP-Ku70a and GFP-Ku80c in vegetative and autogamous cells. In vegetative cells, strong GFP fluorescence is observed in digestive vacuoles (*), as shown for GFP-Ku70a. In autogamous cells, developing MACs are indicated by white arrows, the other DAPI-stained nuclei are fragments from the old vegetative MAC.

B. Colocalization of Pgm-RFP and GFP-Ku80c within developing MACs. Two different developing MACs are shown.

Figure 3: RNAi screen for essential *KU* genes during autogamy

A. Survival of the post-autogamous progeny of cells submitted to different combinations of RNAi. ICL7: RNAi against *ICL7* (Gogendeau et al., 2008), a nonessential gene that encodes an infraciliary lattice centrin. For each condition (RNAi constructs or standard *Klebsiella pneumoniae* (Kp) medium), 30 to 150 post-autogamous cells were analysed. Each bar represents the percentage of viable post-autogamous cells carrying a functional new MAC, for each condition. Error bars represent the Wilson score intervals (95% confidence level), which are appropriate for a small number of trials or for values close to an extreme probability.

B. Visualization of DAPI-stained developing MACs of autogamous cells in *ND7* (control), *KU80c*, *XRCC4/LigIV* and *LigIV + KU80c* RNAi. Developing MACs are indicated by white arrows.

C. Northern blot hybridization of total RNA during a control time course experiment (*ND7* KD) and in a *KU80c* KD, using the *KU80c* probe. V: vegetative cells; T0: 60% of cells with fragmented MAC; the other time-points refer to hours following T0 (Figure S1B for details).

Figure 4: Molecular analysis of genome rearrangements in a *KU80c* KD

A. Detection of DNA fragments with excised or non-excised IES 51G4404 from the surface antigen G^{51} gene. Total genomic DNA was extracted during autogamy of 51ΔA cells submitted to a RNAi against *ICL7* (control) or *KU80c*. To compare the two time-courses, similar autogamy stages were numbered from 1 to 6, based on the observation of DAPI-stained cells (Figure S1C). *Pst*I-hydrolyzed total genomic DNA was run on 1% agarose gels. Southern blots were hybridized with the Gmac probe (in green), which hybridizes to the MAC flanking DNA downstream of the IES.

B. Detection of fragmented or non-fragmented DNA downstream of the G^{51} gene by Southern blot hybridization of *Pst*I-hydrolyzed total genomic DNA (same samples as in A) run on 0.8% agarose gels. The subtelomeric tel51G probe is shown in green. The white box in the bottom right diagram represents telomeric repeats.

C. PCR detection of *de novo* IES excision junctions during autogamy of 51Δ A cells, in an *ICL7* (control) or a *KU80c* KD.

D. PCR detection of IES circle joints during autogamy in an *ICL7* (control) or a *KU80c* KD. Black triangles in C and D represent PCR primers (see Table S1).

Figure 5: Detection of programmed DSBs at IES boundaries during autogamy

A. LMPCR detection of DSBs at MAC or IES ends during autogamy of 51Δ A cells submitted to RNAi against *ICL7* (control) or *KU80c* (same samples as in Figure 4). A DNA Sanger sequencing ladder was used as a size marker. On the diagram, the LMPCR linker is drawn as grey boxes and the *Paramecium* DNA as black lines, with a black dot representing the 3' end generated by Pgm-dependent cleavage. In the *KU80C* KD, the LMPCR signals at 51G4404 MAC ends are likely due to background DNA breaks at the MAC G^{51} locus. This background is absent for IESs of the A^{51} gene, which is absent from the old MAC.

B. TdT tailing of free 3'OH ends during autogamy of 5Δ *A* cells submitted to RNAi against *ICL7* (control) or *KU80c* (same samples as in B). On the diagram, the potentially resected 5' end is represented by a dotted line.

Figure 6: Functional and physical interaction between Pgm and Ku

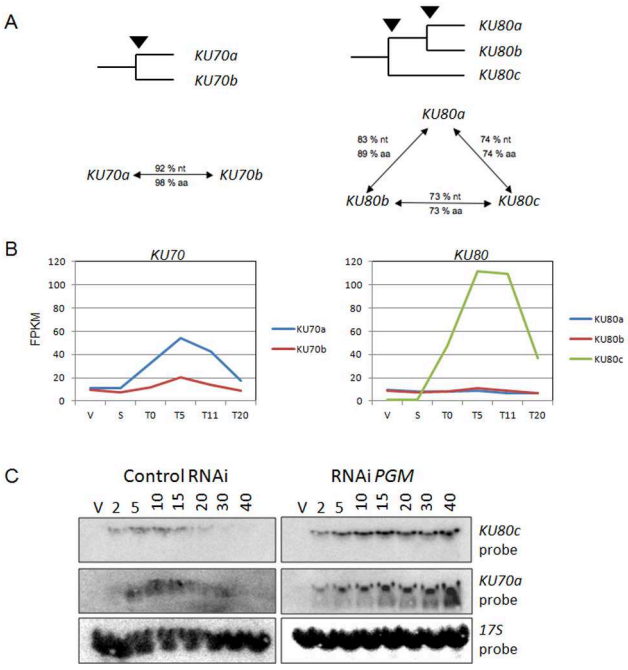
A. Localization of Pgm-GFP during autogamy in transformed cells submitted to *ICL7* (control) or *KU80c* RNAi. To compare the intensities of GFP fluorescence, acquired signals were normalized using the ImageJ software (National Institute of Health). A and E: vegetative cells; B and F: autogamous cells at day 2 of starvation; C and G: day 3; D and H: day 4. Developing MACs are indicated by white arrowheads.

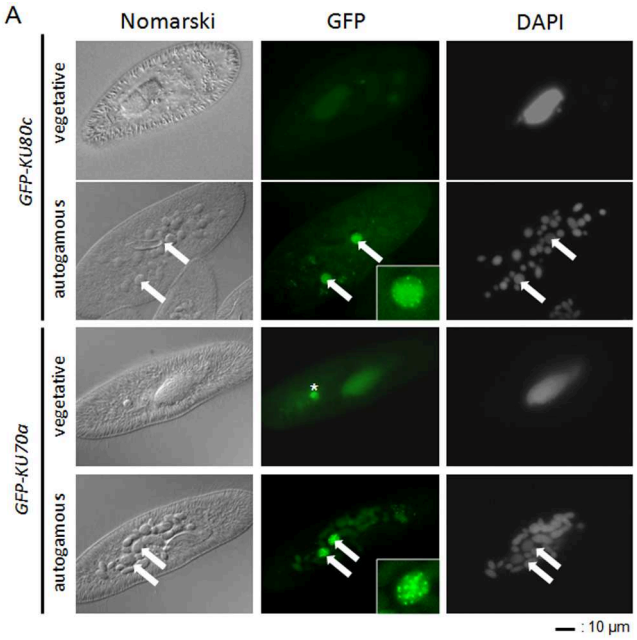
B. Detection of *PGM* mRNA during autogamy, in a control time-course experiment (RNAi against *ND7*) and in a *KU80c* KD. The northern blot shown in Figure 3C was dehybridized and hybridized with the *PGM* probe.

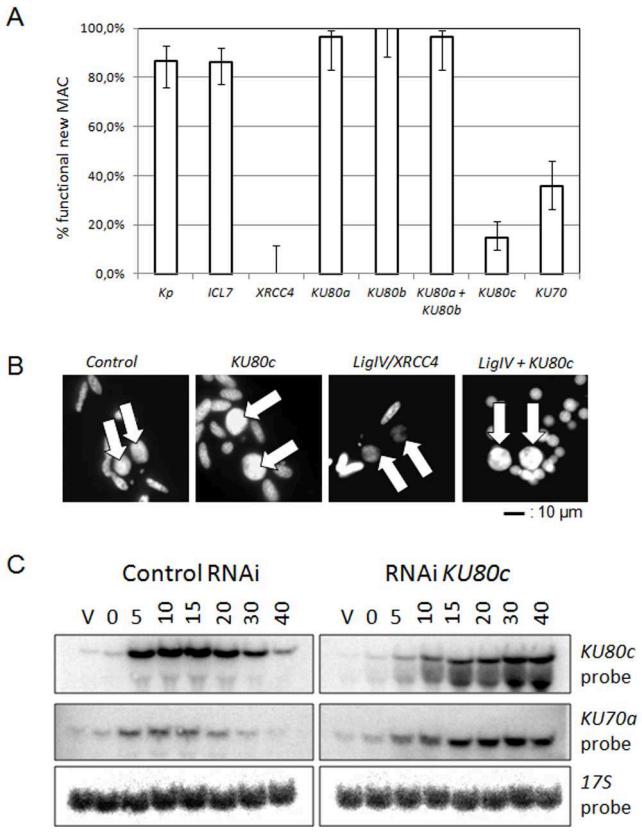
C. Co-precipitation of HA-Ku80c and Ku70a-HA with MBP-Pgm from insect cell extracts (top panel). The input proteins before affinity purification on amylose beads are shown in the bottom panel.

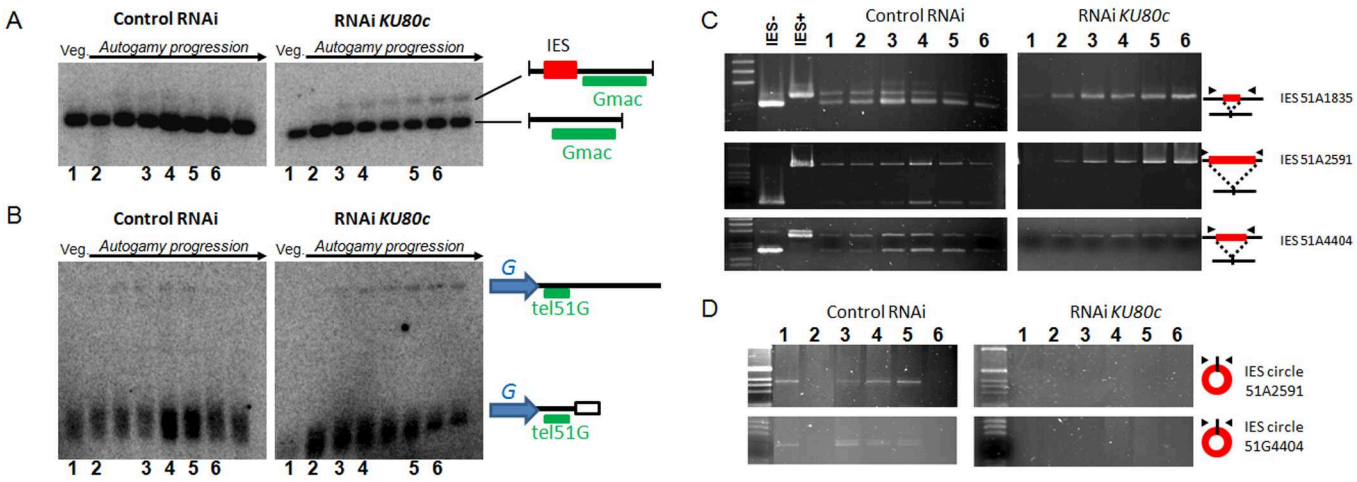
Figure 7: Model for the association of Ku and Pgm in the IES cleavage complex.

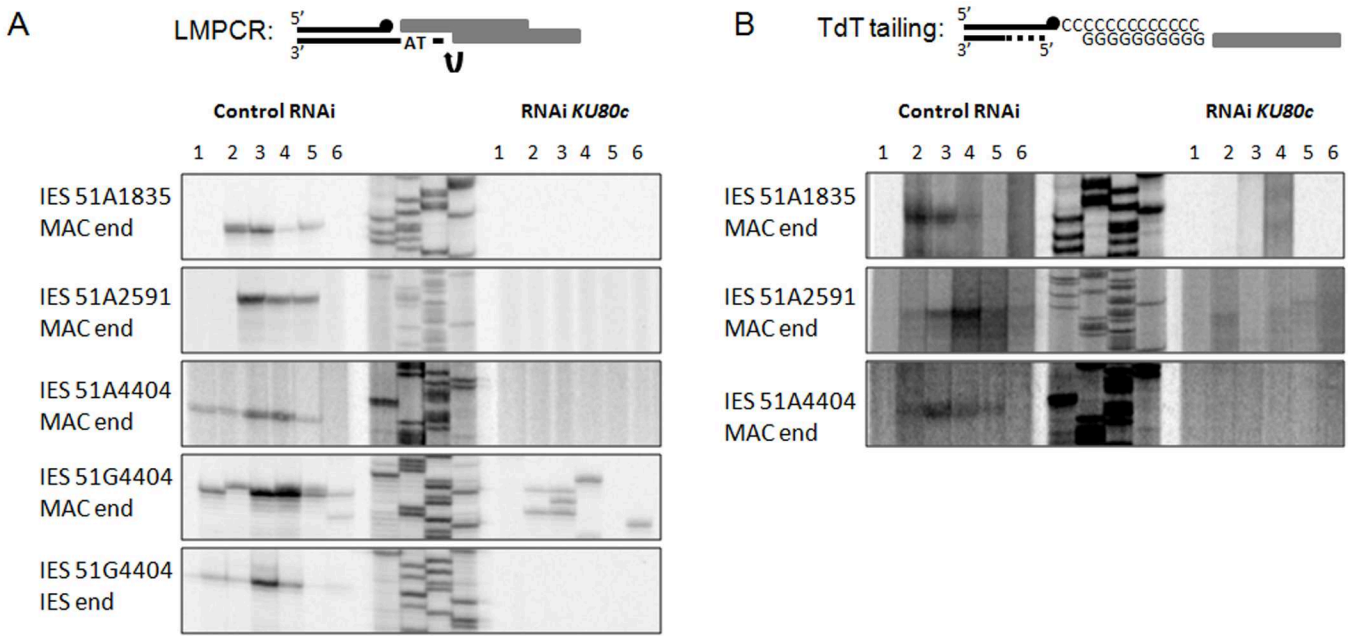
We propose that Pgm and putative other partners (in grey) recognize eliminated regions and bind to IES boundaries. On top, the Ku heterodimer (blue), associated with DNA-PKcs (yellow), activates DNA cleavage by Pgm (symbolized by a switch from a rectangular to an oval grey box in step 2). Following DSB introduction, remodeling of the complex would position Ku on broken DNA ends and allows it to perform its classical role in C-NHEJ-mediated repair (step3). In the absence of Ku (bottom), Pgm would not be activated for DNA cleavage. MAC DNA is represented in black, IES DNA in red.

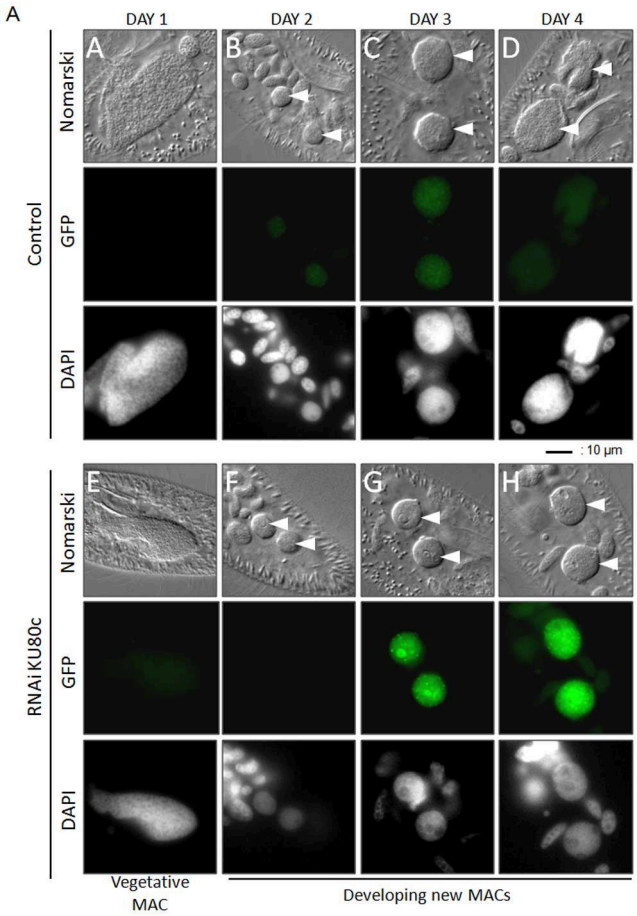

A
KU70a
KU70b
KU80a
KU80b
KU80c
92 % nt
98 % aa
83 % nt
89 % aa
74 % nt
74 % aa
73 % nt
73 % aa
B
KU70
KU80
FPKM
120
100
80
60
40
20
0
V
S
T0
T5
T11
T20
KU70a
KU70b
KU80a
KU80b
KU80c
C
Control RNAi
RNAi PGM
V 2 5 10 15 20 30 40
KU80c probe
KU70a probe
17S probe

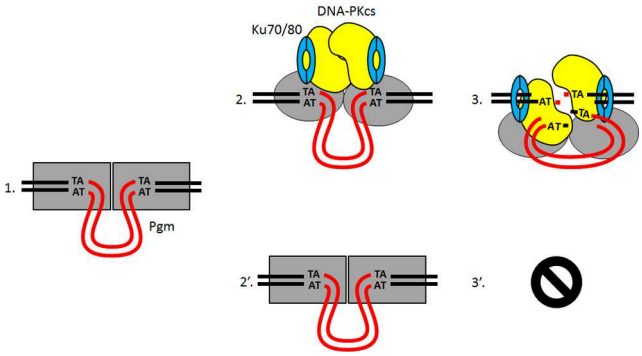

A
Nomarski
GFP
DAPI
GFP-KU80c
vegetative
autogamous
GFP-KU70a
vegetative
autogamous
: 10 μm
B
Nomarski
RFP
GFP
Merge
PGM-RFP
GFP-KU80c
: 5 μm


A
100,0%
80,0%
60,0%
40,0%
20,0%
0,0%
% functional new MAC
Kp
ICL7
XRCC4
KU80a
KU80b
KU80a + KU80b
KU80c
KU70
B
Control
KU80c
LigIV/XRCC4
LigIV + KU80c
: 10 µm
C
Control RNAi
RNAi KU80c
V 0 5 10 15 20 30 40
V 0 5 10 15 20 30 40
KU80c probe
KU70a probe
17S probe


A
Control RNAi
RNAi KU80c
Veg.
Autogamy progression
IES
Gmac
Gmac
1 2 3 4 5 6
1 2 3 4 5 6
B
Control RNAi
RNAi KU80c
Veg.
Autogamy progression
G
tel51G
G
tel51G
1 2 3 4 5 6
1 2 3 4 5 6
C
IES-
IES+
Control RNAi
RNAi KU80c
1 2 3 4 5 6
1 2 3 4 5 6
IES 51A1835
IES 51A2591
IES 51A4404
D
Control RNAi
RNAi KU80c
1 2 3 4 5 6
1 2 3 4 5 6
IES circle 51A2591
IES circle 51G4404


A
LMPCR:
5′
3′
AT
B
TdT tailing:
CCCCCCCCCCCCC
GGGGGGGGGG
Control RNAi
RNAi KU80c
1 2 3 4 5 6
IES 51A1835
MAC end
IES 51A2591
MAC end
IES 51A4404
MAC end
IES 51G4404
MAC end
IES 51G4404
IES end


A
DAY 1
DAY 2
DAY 3
DAY 4
Control
RNAi KU80c
Nomarski
GFP
DAPI
A
B
C
D
E
F
G
H
: 10 µm
Vegetative MAC
Developing new MACs
B
Control RNAi
RNAi KU80c
V 0 5 10 15 20 30 40
PGM probe
17S probe
C
MBP-Pgm + - + - + - +
HA-Ku80c - - - + + + +
HA-Ku70a - + + - - + +
Co-purif.
Input
Ku80c
Ku70a
Antibody α-HA


DNA-PKcs
Ku70/80
1.
TA
AT
TA
AT
Pgm
2.
TA
AT
TA
AT
3.
TA
AT
TA
AT
2'.
TA
AT
TA
AT
3'.

SUPPLEMENTAL INFORMATION TO:

Ku-mediated coupling of DSB introduction and repair during programmed genome rearrangements

A. Marmignon et al.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

RNAi constructs

pXRCC4-R: targets RNAi against *XRCC4* (Kapusta et al., 2011).
pL4440-KU70a-1: a 300-bp fragment from gene *KU70a* (bp 514-813 from ATG start codon) was amplified by PCR using primers ku70A-BamHI_1 and ku70A-KpnI_1, digested with *Bam*HI and *Kpn*I, and ligated between the *Bgl*II and *Kpn*I sites of plasmid L4440 (Kamath et al., 2001).
pL4440-KU80a-1: a 233-bp fragment from gene *KU80a* (bp 40-272 from ATG start codon) was amplified by PCR using primers ku80A-BamHI_1 and ku80A-KpnI_1, digested with *Bam*HI and *Kpn*I, and ligated between the *Bgl*II and *Kpn*I sites of plasmid L4440.
pL4440-KU80a-2: a 450-bp fragment from gene *KU80a* (bp 580-1029 bp from ATG start codon) was amplified by PCR using primers OMB223 and OMB224, digested with *Spe*I and ligated into the *Xba*I site of plasmid L4440.
pL4440-KU80b-1: a 233-bp fragment from gene *KU80b* (bp 40-272 from ATG start codon) was amplified by PCR using primers OMB219 and OMB220, digested with *Spe*I and ligated into the *Xba*I site of plasmid L4440.
pL4440-KU80b-2: a 450-bp fragment from gene *KU80b* (bp 580-1029 bp from ATG start codon) was amplified by PCR using primers OMB221 and OMB222, digested with *Spe*I and ligated into the *Xba*I site of plasmid L4440.
pL4440-KU80c-1: a 206-bp fragment from gene *KU80c* (bp 1021-1226 from ATG start codon) was amplified by PCR using primers ku80C-BamHI_1 and ku80C-KpnI_1bis, digested with *Bam*HI and *Kpn*I, and ligated between the *Bgl*II and *Kpn*I sites of plasmid L4440.
pL4440-KU80c-2: a fragment from gene *KU80C* was amplified by PCR using primers OMB225 and ku80C-KpnI_1bis. Following restriction with *Spe*I, a 450-bp sub-fragment (bp 557-1006 from ATG start codon) was inserted into the *Xba*I site of plasmid L4440.

Construction of GFP and RFP fusions

For the construction of in-frame *GFP-KU* fusions, a GFP-coding fragment adapted to *Paramecium* codon usage (Nowacki et al. 2005) was added by PCR fusion to the 5' end of the *KU70a* or *KU80c* genes. Each construct was inserted in a pUC18 plasmid between the *SphI* and *SacI* sites. As a result, the GFP is fused to the N-terminus of Ku70a and Ku80c and the

fusion proteins are expressed under the control of the *KU70a* and *KU80c* transcription signals (promoters and 3'UTR), respectively.

For the *PGM-RFP* fusion, a *Cla*I site was inserted immediately before the *PGM* STOP codon and a RFP-coding fragment was cloned into this restriction site. All plasmid sequences are available upon request

Preparation of soluble protein extracts from High Five™ cells infected by baculovirus vectors

High Five™ cells were grown in the EX-CELL 405 synthetic medium (Sigma Aldrich) supplemented with Penicillin and Streptomycin. For co-expression experiments, 10^6 cells are infected with *MBP*, *MBP-PGM*, *KU70a-HA* and/or *HA-KU80c* recombinant baculoviruses. At 48 hrs post-infection, cells were collected, washed with cold PBS buffer and re-suspended in 500µl of lysis buffer containing 10mM Tris-HCL pH 7.5, 150mM NaCl, 1% Triton X-100(v/v), 1mM EDTA, 1mM DTT and 0.5% NaDesoxycholate (m/v) in the presence of protease inhibitors (Complete Protease Inhibitor Cocktail Tablets, Roche). Cells were lysed for 20 min on a rotating wheel, then centrifuged at 4°C for 15 min at 10,000 g.

LEGENDS TO SUPPLEMENTARY FIGURES AND TABLES

Figure S1: Progression of autogamy in the cultures used in this study

For each time-point, cell stages were monitored by DAPI staining. Veg: vegetative cells. Skeins: cells with elongating old MAC at the beginning of fragmentation. Meiosis*: cells with detectable meiotic MICs (which is an underestimate of the actual fraction of meiotic cells). Fragments: cells with fully fragmented old MAC. 2 anlagen: cells with two visible new developing MACs. postA: post autogamous cells with one new MAC and a few fragments of the old MAC.

A. Autogamy time course of strain 51 submitted to RNAi against *ND7* or *PGM*. The ND7 and PGM-1 constructs used for RNAi were described in (Baudry et al., 2009).

B. Autogamy time course of strain 51 submitted to RNAi against *ND7* or *KU80c* (using the KU80c-2 construct, see Figure S3).

C. Autogamy time course of strain 51ΔA submitted to RNAi against *ICL7* (the ICL7a RNAi construct was described in (Gogendeau et al., 2008)) or *KU80c* (RNAi construct KU80c-2).

Figure S2: Co-localization of Pgm GFP and RFP fusions during autogamy

The PGM-GFP and PGM-RFP constructs were micro-injected together into the vegetative MAC of 51 *nd7-1* cells. Transformants were allowed to undergo autogamy in standard growth medium. Each panel displays the same developing MAC from one autogamous cell.

Figure S3: Maps of the *KU70* and *KU80* genes of *P. tetraurelia*

The maps show the position of the inserts used for RNAi constructs (blue) and hybridization probes (red, see Table S1). For *KU70* genes, all fragments were designed from the *KU70a* sequence. Specific fragments were designed for each *KU80* gene.

Figure S4: Detection of *de novo* IES excision junctions in different RNAi conditions

The excision of IESs 51A1835, 51A2591 and 51A4404 during autogamy of strain 51ΔA (4 days of starvation) was tested by PCR amplification. RNAi conditions are indicated next to each panel. Here, starvation was prolonged for one additional day relative to the other experiments shown in the paper. Therefore, day 4 would correspond to approximately 20 hours following time-point 6 in Figure 4. The low levels of IES- forms were detected very late during autogamy might be attributable to a decrease in the efficiency of RNAi upon

prolonged starvation, as discussed in the main text. All PCR primers are displayed in Table S1.

Figure S5: Control co-precipitation experiments for the study of the physical interaction between Ku and Pgm

A. The interaction between MBP-Pgm and Ku is mediated by the Pgm protein sequence. MBP and MBP-Pgm were co-expressed with Ku70a-HA or HA-Ku80c in insect cells using recombinant baculovirus vectors. MBP was purified from full cell extracts using amylose-coupled magnetic beads and co-purification of HA tagged proteins was tested on western blots. A low level of non-specific interaction of Ku70a with the amylose beads is observed when Ku70a is expressed alone, or together with the MBP. Significant enrichment in Ku70a-HA and HA-Ku80c is observed only when each protein is co-expressed with Pgm (Top left panel). Expression and precipitation of the MBP and the MBP-Pgm protein were checked by western blot (anti-MBP-HRP coupled antibodies, NewEngland Biolabs) on the same membrane after harsh stripping (Bottom left panel). The expression of Ku70a-HA and HA-Ku80c proteins in the input cell extracts was checked by western blot (Top right panel).

B. DNaseI treatment does not abolish the Ku-Pgm interaction. The co-precipitation experiment was performed as previously described, using MBP-Pgm, HA-Ku80c and 6HIS-Ku70a recombinant proteins produced from baculovirus vectors. EDTA was removed from the lysis buffer and replaced by 10mM $MgCl_2$. Half of the sample was treated with DNaseI (10µg) during the 2-hr incubation with the amylose-coupled magnetic beads, then the samples were treated as described previously. The interaction between HA-Ku80c and Pgm is preserved after DNaseI treatment.

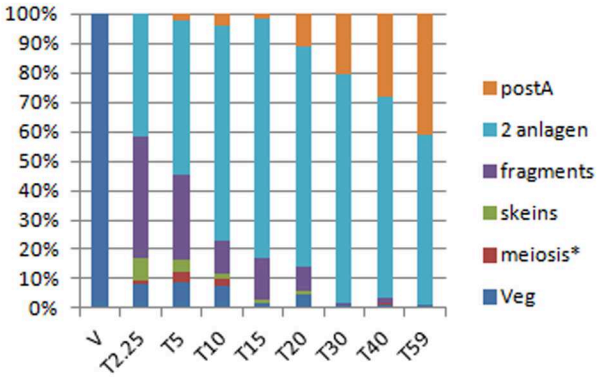
Table S1: oligonucleotides used in this study

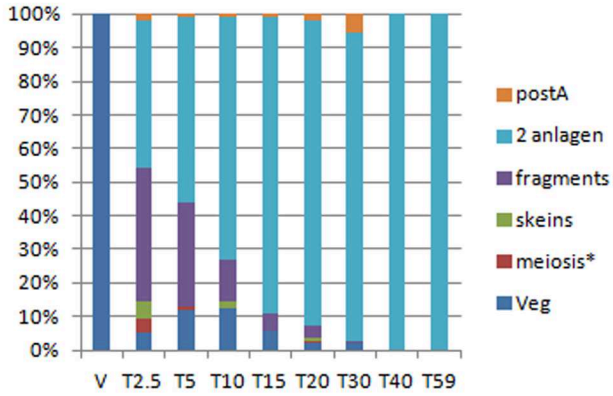
A

RNAi *PGM* strain 51

ND7	V	T2.25	T5	T10	T15	T20	T30	T40	T59
Veg	100	9	9	8	2	5	1	1	1
meiosis*	0	1	4	3	0	0	0	1	0
skeins	0	8	4	2	1	1	0	0	0
fragments	0	44	30	12	16	9	1	2	0
2 anlagen	0	44	55	80	92	80	89	72	60
postA	0	0	2	4	2	12	23	30	43
	100	106	104	109	113	107	114	106	104

PGM	V	T2.5	T5	T10	T15	T20	T30	T40	T59
Veg	100	6	12	13	6	2	2	0	0
meiosis*	0	5	1	0	0	1	0	0	0
skeins	0	6	0	2	0	1	0	0	0
fragments	0	46	32	13	5	4	1	0	0
2 anlagen	0	51	56	75	90	100	100	100	100
postA	0	2	1	1	1	2	6	0	0
	100	116	102	104	102	110	109	100	100



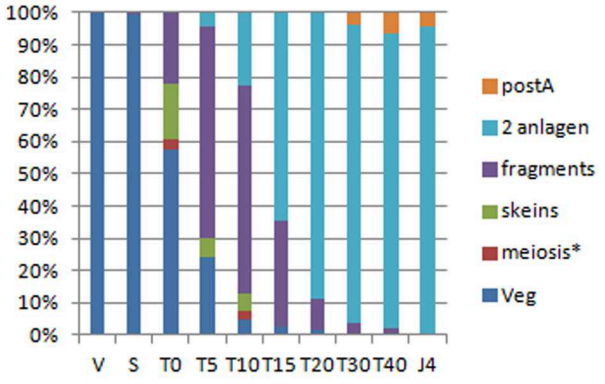


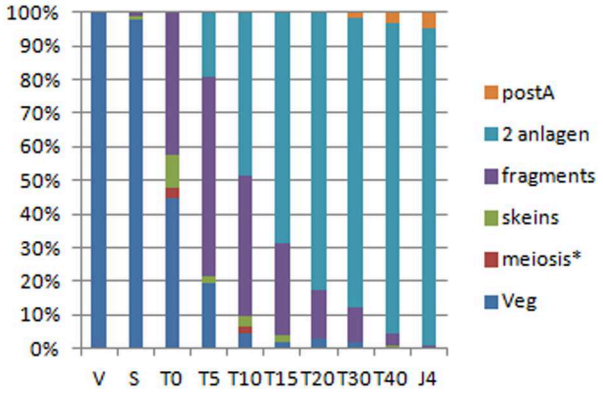
B

RNAi *KU80c* strain 51

ND7	V	S	T0	T5	T10	T15	T20	T30	T40	J4
Veg	100	198	54	16	4	2	1	0	0	0
meiosis*	0	0	3	0	2	0	0	0	0	0
Skeins	0	0	16	4	4	0	0	0	0	0
Fragments	0	1	21	44	51	23	7	3	2	0
2 anlagen	0	0	0	3	18	46	62	72	84	107
postA	0	0	0	0	0	0	0	3	6	5
	100	199	94	67	79	71	70	78	92	112

KU80C	V	S	T0	T5	T10	T15	T20	T30	T40	J4
Veg	100	85	42	17	4	2	3	2	0	0
meiosis*	0	0	3	0	2	0	0	0	0	0
skeins	0	1	9	2	3	2	0	0	1	0
fragments	0	1	40	53	40	30	15	11	4	1
2 anlagen	0	0	0	17	46	74	87	92	112	141
postA	0	0	0	0	0	0	0	2	4	7
	100	87	94	89	95	108	105	107	121	149



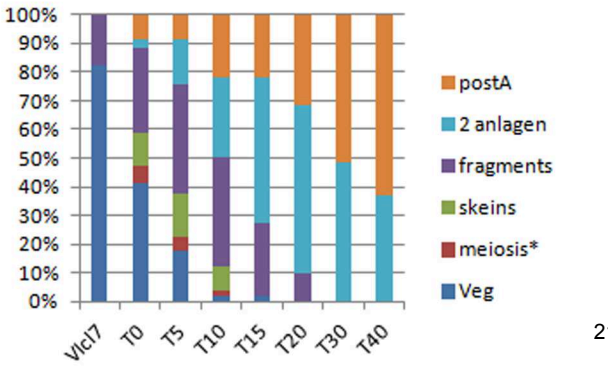


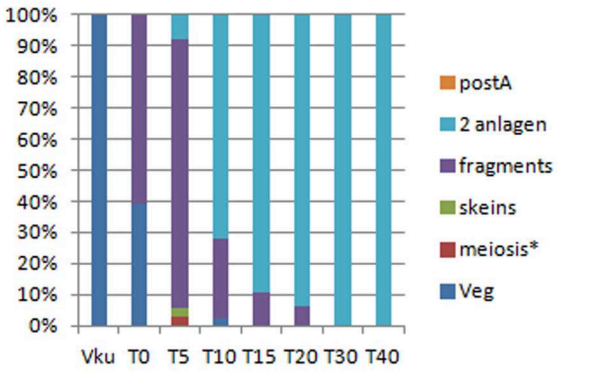
C

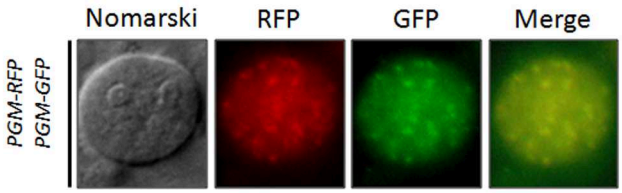
RNAi *KU80C* strain 51ΔA

ICL7	VIcl7	T0	T5	T10	T15	T20	T30	T40
Veg	82	14	8	1	1	0	0	0
meiosis*	0	2	2	1	0	0	0	0
skeins	0	4	7	4	0	0	0	0
fragments	18	10	17	19	15	4	0	0
2 anlagen	0	1	7	14	30	24	18	10
postA	0	3	4	11	13	13	19	17
	100	34	45	50	59	41	37	27

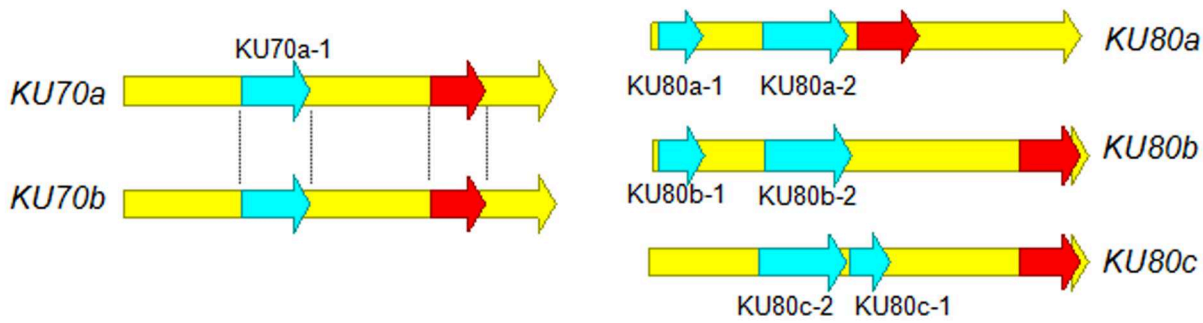
KU80C	Vku	T0	T5	T10	T15	T20	T30	T40
Veg	100	39	0	1	0	0	0	0
meiosis*	0	0	1	0	0	0	0	0
skeins	0	0	1	0	0	0	0	0
fragments	0	61	32	11	6	3	0	0
2 anlagen	0	0	3	31	51	45	69	41
postA	0	0	0	0	0	0	0	0
	100	100	37	43	57	48	69	41

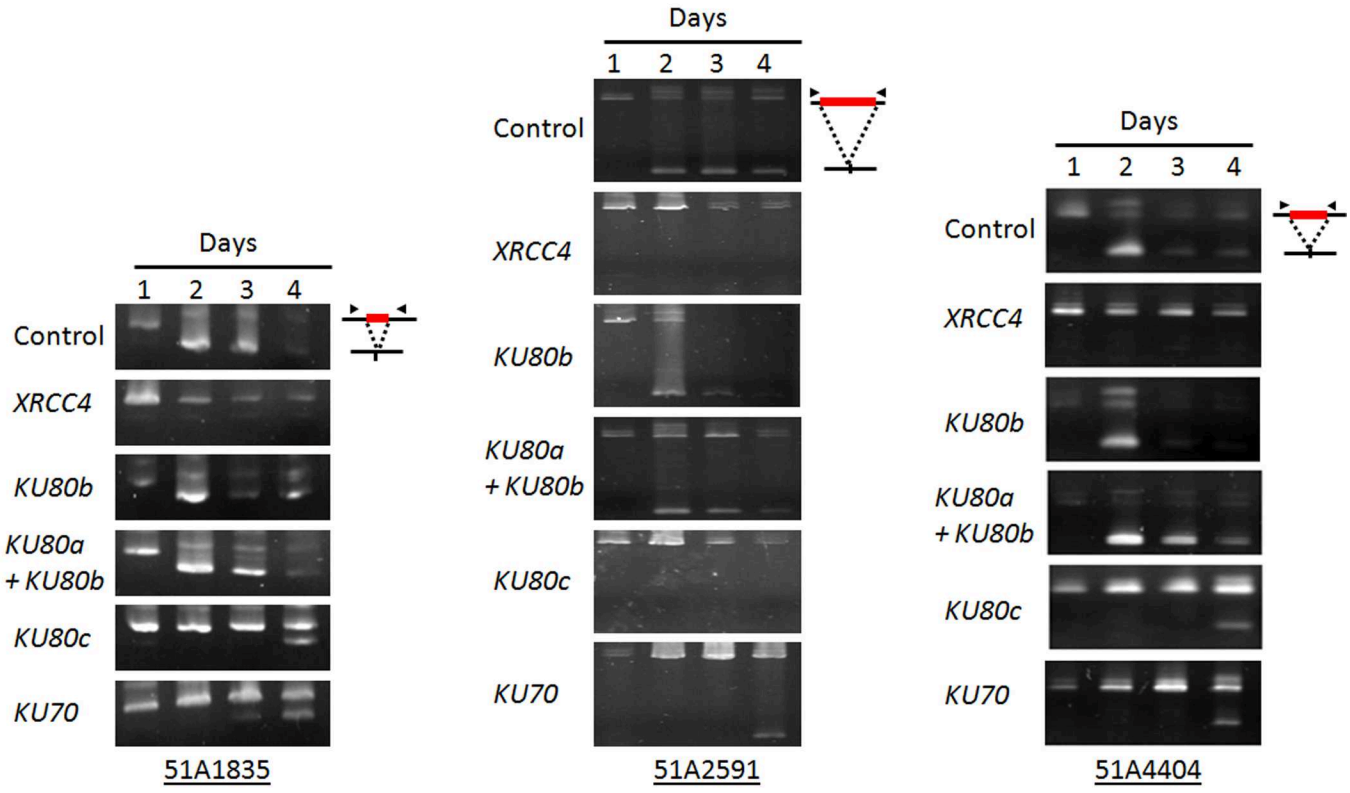





Nomarski
RFP
GFP
Merge
PGM-RFP
PGM-GFP

A


KU70a-1
KU70a
KU70b
KU80a
KU80a-1
KU80a-2
KU80b
KU80b-1
KU80b-2
KU80c
KU80c-2
KU80c-1


Days
1 2 3 4
Control
XRCC4
KU80b
KU80a
+ KU80b
KU80c
KU70
51A1835
Days
1 2 3 4
Control
XRCC4
KU80b
KU80a
+ KU80b
KU80c
KU70
51A2591
Days
1 2 3 4
Control
XRCC4
KU80b
KU80a
+ KU80b
KU80c
KU70
51A4404

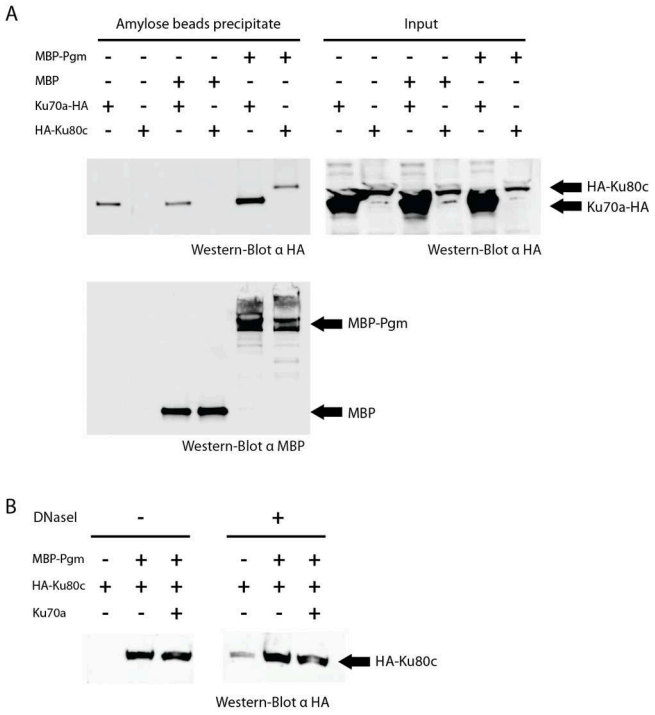

A
Amylose beads precipitate
Input
MBP-Pgm
MBP
Ku70a-HA
HA-Ku80c
HA-Ku80c
Ku70a-HA
Western-Blot α HA
Western-Blot α HA
MBP-Pgm
MBP
Western-Blot α MBP
B
DNaseI
MBP-Pgm
HA-Ku80c
Ku70a
HA-Ku80c
Western-Blot α HA

Table S1: Oligonucleotides used in this study

name	sequence (5' to 3')	use
ku70A-BamHI_1	gaagacaggatccAGATAAGAACAAATTCAATATGCGTA	PCR feeding insert KU70
ku70A-KpnI_1	gaagacaggtaccTTCAAAGCTCTTTTCTTAAATTCCT	PCR feeding insert KU70
ku80A-BamHI_1	gaagacaggatccGGTGCCTCAATGTATGAACCATACA	PCR feeding insert KU80a-1
ku80A-KpnI_1	gaagacaggtaccGGTAATTCTGTCAGATTTCTATAAAAC	PCR feeding insert KU80a-1
OMB223	ggactagTACGAAAGTTAAGAGCAATCAATCAATTC	PCR feeding insert KU80a-2
OMB224	ggactagtCAATAATTAAAATGACCTAACACATTAATAC	PCR feeding insert KU80a-2
OMB219	ggactagtGGTGCCTCAATGTATGAACCATACAAGTAG	PCR feeding insert KU80b-1
OMB220	ggactagtGGTAATTCTGTCAGATTTCTATAAAACTC	PCR feeding insert KU80b-1
OMB221	ggactagTACTAACATTAAGAACAATCAATAAATTC	PCR feeding insert KU80b-2
OMB222	ggactagtAAGCAATTAAAATGATCTATTACATTAATAC	PCR feeding insert KU80b-2
ku80C-BamHI_1	gaagacaggatccTAATTATTAGGCTTTGTAGATCGATC	PCR feeding insert KU80c-1
ku80C-KpnI_1bis	gaagacaggtaccTGGGGTAATAACATAATCAATTTAGGT	PCR feeding inserts KU80c-1 & KU80c-2
OMB225	ggactagtATAATAGAATGCTTACAGCTTCATATCAATG	PCR feeding insert KU80c-2
OMB063	agacaagtagggaatccacttctagtaatc	PCR around IES 51A1835
OMB064	taatgtattgataaggcttgctctacagcc	PCR around IES 51A1835
OMB068	acaccaagcgaaacatgcacagtcg	PCR around IES 51A2591
OMB256	GATGTAGCATAACATTTATCAACAATCCAT	PCR around IES 51A2591
OMB069	ccagttattgaactgcaacttactgcagtg	PCR around IES 51A4404
OMB097	TAAATGTTCAGCTTACAACGCAGCT	PCR around IES 51A4404
OMB365	AGATTTATATCTTTTTTCTCAAATTCAGC	IES circle 51A2591
OMB184	CAATATTATACATCTAGAACTTATAGTTAG	IES circle 51A2591
OMB181	TTTTGAAATATTTTCAAGTTTTTGGACTAC	IES circle 51G4404
OMB182	ACAATATATATTTACTTGATAATATTTTCC	IES circle 51G4404
OMB062	gtagtacaagatttttcgacacaagttgag	LMPCR/TdT 51A1835 MAC end
OMB065	ggttgcgtaacacttcctcttaaatgtgag	LMPCR/TdT 51A1835 MAC end
OMB066	gaagtctaatggataaccttgtggatggac	LMPCR/TdT 51A1835 MAC end
OMB145	AATTGTAAATTGACTTCAGCAAATAAAAAA	LMPCR/TdT 51A2591 MAC end
OMB212	ATGTGTTTGGACTGGATTGGCATGTAGAAG	LMPCR/TdT 51A2591 MAC end
OMB215	AGTTCCTTTGAAAGATGTGCAAGCTCCAGA	LMPCR/TdT 51A2591 MAC end
OMB069	ccagttattgaactgcaacttactgcagtg	LMPCR/TdT 51A4404 MAC end
OMB070	tggaatagtgctgcatcaccagctgcttgc	LMPCR/TdT 51A4404 MAC end
OMB226	ACCAGCTGCTTGCATTCAAATATCCACAGT	LMPCR/TdT 51A4404 MAC end
OMB113	TGCATATGTTACTGGAACTGGATTGGTAGC	LMPCR 51G4404 MAC end

OMB114	ACTGTTGCTACACATTGTGCATATGTTACT	LMPCR 51G4404 MAC end
OMB213	GCTGTAAGATTAACATTGAGCATGATCAAG	LMPCR 51G4404 MAC end
OMB300	AAAGGCTAATTTGGATGAATGAGCATTAAATC	LMPCR 51G4404 IES end
OMB301	GGACTACTTTTGAAATTGAATTATAACAAAGGC	LMPCR 51G4404 IES end
OMB056	gaattcggatccgctcggaccgtggc	LMPCR
OMB032	gctcggaccgtggctagcattagtgagtgggggggggggggggg	TdT tailing
OMB176	ACCCGTGACTGCCATGGTAGTCCAATACA	17S probe
OMB235	GTCTAGTGTGGACATGGTTGTAGCTATTGA	KU80a probe
OMB234	TCTGGTTCTCCAGATTCGTTTGGAAGTGCT	KU80a probe
OMB233	GTTATTAATGCACTAGTTGATTAATCAGTG	KU80b probe
OMB232	TTTGTAATTAGTCCACCATTTACTAATGTTTAT	KU80b probe
OMB231	GTCATTATTTAATTTGTGGATCAATCTCTC	KU80c probe
OMB230	GTTATAATTGATCAACTAATTATTTATGACTAA	KU80c probe
OMB237	TATGGCAAACCCTTAGATAGGAGATACAAC	KU70 probe
OMB236	TTTCAATATCATTTATTGAATCAATGCATC	KU70 probe

REFERENCES TO SUPPLEMENTARY INFORMATION

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., and Bétermier, M. (2009). PiggyMac, a domesticated *piggyBac* transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. Genes Dev. *23*, 2478-2483.

Gogendeau, D., Klotz, C., Arnaiz, O., Malinowska, A., Dadlez, M., de Loubresse, N.G., Ruiz, F., Koll, F., and Beisson, J. (2008). Functional diversification of centrins and cell morphological complexity. J Cell Sci. *121*, 65-74.

Kamath, R.S., Martinez-Campos, M., Zipperlen, P., Fraser, A.G., and Ahringer, J. (2001). Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in *Caenorhabditis elegans*. Genome Biol *2*, RESEARCH0002.

Kapusta, A., Matsuda, A., Marmignon, A., Ku, M., Silve, A., Meyer, E., Forney, J., Malinsky, S., and Bétermier, M. (2011). Highly precise and developmentally programmed genome assembly in *Paramecium* requires Ligase IV-dependent end joining. PloS Genetics *7*, e1002049.

ACKNOWLEDGMENTS

We would like to thank Nathalie Mathy for expert technical assistance and all students of the 2008 Pasteur course on “Analysis of Genomes” for their contribution to the initial screening for essential C-NHEJ genes in *P. tetraurelia*. This work has benefited from the facilities and expertise of the IMAGIF high-throughput sequencing platform (Centre de Recherche de Gif - www.imagif.cnrs.fr). Special thanks to the French community of *Paramecium* labs for extremely stimulating discussions and to Linda Sperling for critical reading of the manuscript. This work has been supported by core funding from the CNRS and by grants from the Agence Nationale de la Recherche (ANR 2010-BLAN-1603 and ANR-12-BSV6-0017) and the Fondation ARC (#SFI20121205487). AM was supported by PhD fellowships from the Ministère de l’Enseignement Supérieur et de la Recherche and from the Fondation ARC.

REFERENCES

Arnaiz, O., Gout, J.F., Bétermier, M., Bouhouche, K., Cohen, J., Duret, L., Kapusta, A., Meyer, E., and Sperling, L. (2010). Gene expression in a paleopolyploid: a transcriptome resource for the ciliate *Paramecium tetraurelia*. BMC Genomics *11*, 547.

Arnaiz, O., Mathy, N., Baudry, C., Malinsky, S., Aury, J.M., Denby-Wilkes, C., Garnier, O., Labadie, K., Lauderdale, B.E., Le Mouel, A., *et al.* (2012). The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: Evolutionary dynamics of internal eliminated sequences. PloS Genetics *8*, e1002984.

Arnaiz, O., and Sperling, L. (2011). ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate *Paramecium tetraurelia*. Nucleic Acids Res *39*, D632-D636.

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N., *et al.* (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. *444*, 171-178.

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., and Bétermier, M. (2009). PiggyMac, a domesticated *piggyBac* transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. Genes Dev. *23*, 2478-2483.

Beall, E.L., and Rio, D.C. (1996). *Drosophila* IRBP/Ku p70 corresponds to the mutagen-sensitive mus309 gene and is involved in P-element excision *in vivo*. Genes Dev *10*, 921-933.

Bétermier, M. (2004). Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate *Paramecium*. Res Microbiol *155*, 399-408.

Bétermier, M., Duharcourt, S., Seitz, H., and Meyer, E. (2000). Timing of developmentally programmed excision and circularization of *Paramecium* internal eliminated sequences. Mol. Cell. Biol. *20*, 1553-1561.

Borde, V. (2007). The multiple roles of the Mre11 complex for meiotic recombination. Chromosome Res *15*, 551-563.

Chalker, D.L., and Yao, M.C. (2011). DNA elimination in ciliates: transposon domestication and genome surveillance. Annu Rev Genet *45*, 227-246.

Chapman, J.R., Taylor, M.R., and Boulton, S.J. (2012). Playing the end game: DNA double-strand break repair pathway choice. Mol Cell *47*, 497-510.

Cheng, C.Y., Vogt, A., Mochizuki, K., and Yao, M.C. (2010). A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in *Tetrahymena thermophila*. Mol Biol Cell *21*, 1753-1762.

Coyne, R.S., Lhuillier-Akakpo, M., and Duharcourt, S. (2012). RNA-guided DNA rearrangements in ciliates: is the best genome defence a good offence? Biol Cell *104*, 309-325.

Dubois, E., Bischerour, J., Marmignon, A., Mathy, N., Régnier, V., and Bétermier, M. (2012). Transposon Invasion of the *Paramecium* Germline Genome Countered by a Domesticated PiggyBac Transposase and the NHEJ Pathway. Int J Evol Biol *2012*, 436196.

Duharcourt, S., Butler, A., and Meyer, E. (1995). Epigenetic self-regulation of developmental excision of an internal eliminated sequence in *Paramecium tetraurelia*. Genes Dev. *9*, 2065-2077.

Fass, J.N., Joshi, N.A., Couvillion, M.T., Bowen, J., Gorovsky, M.A., Hamilton, E.P., Orias, E., Hong, K., Coyne, R.S., Eisen, J.A., *et al.* (2011). Genome-Scale Analysis of Programmed DNA Elimination Sites in Tetrahymena thermophila. G3 (Bethesda) *1*, 515-522.

Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet *41*, 331-368.

Galvani, A., and Sperling, L. (2002). RNA interference by feeding in *Paramecium*. Trends Genet *18*, 11-12.
Garnier, O., Serrano, V., Duharcourt, S., and Meyer, E. (2004). RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*. Mol. Cell. Biol. *24*, 7370-7379.
Gogendeau, D., Klotz, C., Arnaiz, O., Malinowska, A., Dadlez, M., de Loubresse, N.G., Ruiz, F., Koll, F., and Beisson, J. (2008). Functional diversification of centrins and cell morphological complexity. J Cell Sci. *121*, 65-74.
Gratias, A., and Bétermier, M. (2003). Processing of double-strand breaks is involved in the precise excision of *Paramecium* IESs. Mol. Cell. Biol. *23*, 7152-7162.
Gratias, A., Lepère, G., Garnier, O., Rosa, S., Duharcourt, S., Malinsky, S., Meyer, E., and Bétermier, M. (2008). Developmentally programmed DNA splicing in *Paramecium* reveals short-distance crosstalk between DNA cleavage sites. Nucleic Acids Res *36*, 3244-3251.
Izsvak, Z., Stuwe, E.E., Fiedler, D., Katzer, A., Jeggo, P.A., and Ivics, Z. (2004). Healing the wounds inflicted by sleeping beauty transposition by double-strand break repair in mammalian somatic cells. Mol Cell. *13*, 279-290.
Kapusta, A., Matsuda, A., Marmignon, A., Ku, M., Silve, A., Meyer, E., Forney, J., Malinsky, S., and Bétermier, M. (2011). Highly precise and developmentally programmed genome assembly in *Paramecium* requires Ligase IV-dependent end joining. PloS Genetics *7*, e1002049.
Karamysheva, Z., Wang, L., Shrode, T., Bednenko, J., Hurley, L.A., and Shippen, D.E. (2003). Developmentally programmed gene elimination in *Euplotes crassus* facilitates a switch in the telomerase catalytic subunit. Cell *113*, 565-576.
Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.
Lepère, G., Bétermier, M., Meyer, E., and Duharcourt, S. (2008). Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in *Paramecium tetraurelia*. Genes Dev. *22*, 1501-1512.
Lepère, G., Nowacki, M., Serrano, V., Gout, J.F., Guglielmi, G., Duharcourt, S., and Meyer, E. (2009). Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*. Nucleic Acids Res. *37*, 903-915.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.
Lieber, M.R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annu Rev Biochem *79*, 181-211.
Lin, I.T., Chao, J.L., and Yao, M.C. (2012). An essential role for the DNA breakage-repair protein Ku80 in programmed DNA rearrangements in *Tetrahymena* thermophila. Mol Biol Cell *23*, 2213-2225.
Longhese, M.P., Bonetti, D., Guerini, I., Manfrini, N., and Clerici, M. (2009). DNA double-strand breaks in meiosis: checking their formation, processing and repair. DNA Repair (Amst). *8*, 1127-1138.
McVey, M., and Lee, S.E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet *24*, 529-538.
Nowak, J.K., Gromadka, R., Juszczuk, M., Jerka-Dziadosz, M., Maliszewska, K., Mucchielli, M.H., Gout, J.F., Arnaiz, O., Agier, N., Tang, T., *et al.* (2011). Functional study of genes essential for autogamy and nuclear reorganization in *Paramecium*. Eukaryot Cell *10*, 363-372.

Raval, P., Kriatchko, A.N., Kumar, S., and Swanson, P.C. (2008). Evidence for Ku70/Ku80 association with full-length RAG1. Nucleic Acids Res *36*, 2060-2072.
Robert, V., and Bessereau, J.L. (2007). Targeted engineering of the *Caenorhabditis elegans* genome following *Mos1*-triggered chromosomal breaks. Embo J *26*, 170-183.
Saveliev, S.V., and Cox, M.M. (1995). Transient DNA breaks associated with programmed genomic deletion events in conjugating cells of *Tetrahymena thermophila*. Genes Dev. *9*, 248-255.
Saveliev, S.V., and Cox, M.M. (1996). Developmentally programmed DNA deletion in *Tetrahymena thermophila* by a transposition-like reaction pathway. EMBO J. *15*, 2858-2869.
Saveliev, S.V., and Cox, M.M. (2001). Product analysis illuminates the final steps of IES deletion in *Tetrahymena thermophila*. EMBO J. *20*, 3251-3261.
Schatz, D.G., and Ji, Y. (2011). Recombination centres and the orchestration of V(D)J recombination. Nat Rev Immunol *11*, 251-263.
Schatz, D.G., and Swanson, P.C. (2011). V(D)J recombination: mechanisms of initiation. Annu Rev Genet *45*, 167-202.
Skouri, F., and Cohen, J. (1997). Genetic approach to regulated exocytosis using functional complementation in *Paramecium*: identification of the *ND7* gene required for membrane fusion. Mol Biol Cell *8*, 1063-1071.
Symington, L.S., and Gautier, J. (2011). Double-strand break end resection and repair pathway choice. Annu Rev Genet *45*, 247-271.
Taylor, J.S., and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet *38*, 615-643.
Yant, S.R., and Kay, M.A. (2003). Nonhomologous-end-joining factors regulate DNA repair fidelity during *Sleeping Beauty* element transposition in mammalian cells. Mol Cell Biol *23*, 8505-8518.

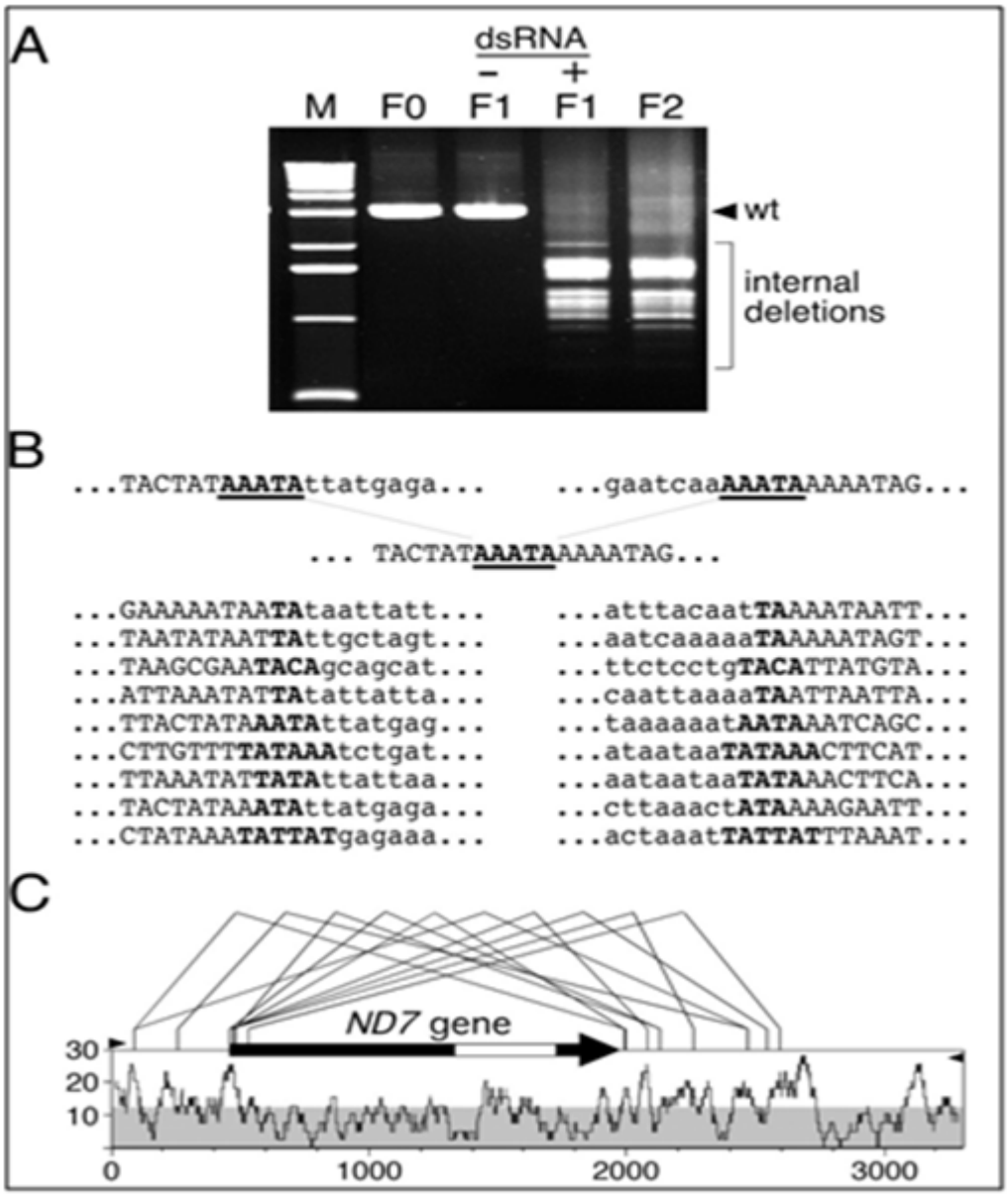


Figure 43 : Caractérisation de la délétion interne du gène ND7. A. Par PCR. B. Par séquençage des produits de jonctions. C. Les délétions sont indiquées sur une carte du gène ND7 (Garnier et al., 2004).

Résultats supplémentaires sur le rôle de Ku dans les réarrangements programmés.

Caractérisation de la délétion induite du gène *ND7*.

Avant d'analyser le rôle de différents gènes *KU* dans les réarrangements programmés imprécis, il me fallait caractériser un site mic dans des conditions normales d'autogamie. Il n'existe pas de site naturel de réarrangement imprécis caractérisé chez *Paramecium tetraurelia* mais le comportement de la délétion imprécise, induite et héritable du gène *ND7* semble assez proche de celui attendu pour des sites naturels, on sait qu'il y a des délétions internes de taille variable (figure 43) mais la fragmentation des chromosomes à ce site est inconnue (Garnier et al., 2004). Dans un premier temps il est donc nécessaire de caractériser la délétion somatique, imprécise, induite et héritable du gène *ND7* dans des souches Δ*ND7* pour estimer la quantité relative de chromosomes fragmentés par rapport aux délétions internes.

Le gène *ND7* est situé sur le chromosome MAC numéro 5 long de 770kb. Comme précisé en introduction, l'élimination imprécise de séquences germinales peut emporter de larges fragments de séquences de plusieurs kb, l'approche moléculaire par PCR est donc ici inadaptée. De plus la PCR ne permettrait d'observer que des produits de délétion interne et non les événements de fragmentation. La technique d'électrophorèse en champ pulsé pourrait permettre d'observer le chromosome MAC entier ou l'apparition de fragments de chromosomes plus petits de l'ordre de 173kb et 597kb correspondants à des produits de fragmentation. Tandis que la méthode d'électrophorèse classique permettrait de détecter des délétions de tailles variables entre deux sites de restriction PstI et SalI distant de 8kb encadrant le gène *ND7*.

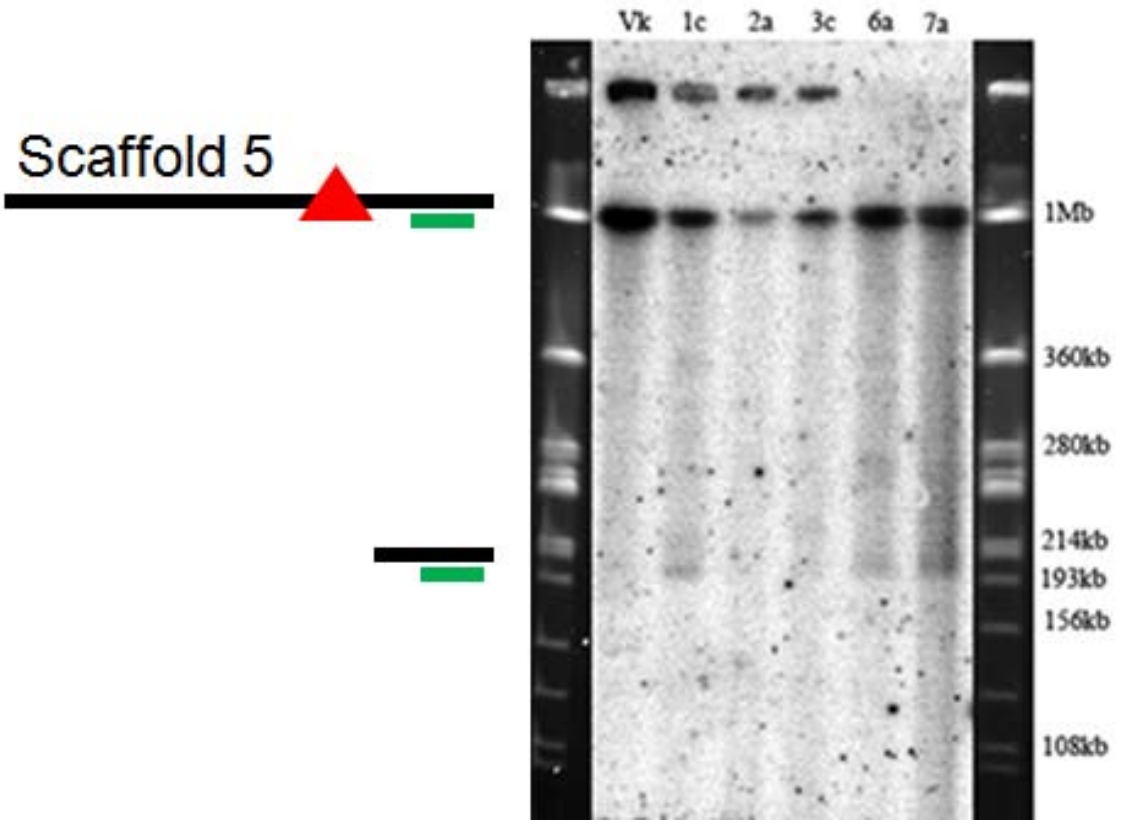


Figure 44 : Gel d'électrophorèse en champ pulsé d'ADN de cellules délétées du gène ND7. Hybridation d'une sonde spécifique du bras droit du scaffold 5. (Conditions de migration : 21h, 60-120 sec, 6V, 120°)

A l'échelle du chromosome entier

De l'ADN de clones délétés ou non du gène *ND7* dans leur MAC a été extrait et mis à migrer sur gel d'agarose par la méthode d'électrophorèse en champ pulsé. Après transfert sur membrane, une sonde marquée spécifique à la séquence complémentaire du « bras droit » (173kb) du chromosome a été hybridée (figure 44). Cette sonde doit révéler une forme longue d'environ 770kb correspondant au chromosome entier ou avec une délétion interne du gène *ND7* d'environ 1,5kb ainsi que le bras droit du chromosome fragmenté, s'il y a lieu. Néanmoins cette méthode ne permet pas de différencier chromosome entier ou porteur d'une délétion interne du gène *ND7*. On remarque que seuls les clones Δ*ND7* présentent un phénotype de fragmentation du chromosome. En effet on observe dans ces pistes une bande aux environs de 200kb qui n'est toutefois pas très nette : il est fort possible qu'il y ait des fragments de taille variable au niveau de cette bande correspondant à une variabilité sur la position de fragmentation et/ou d'addition de télomère au sein d'un même macronoyau. Dans la piste correspondant à un ADN de clone sauvage, on ne remarque pas de bande à cette taille, comme attendu.

Curieusement, dans tous les clones testés, délétés ou non, la même sonde révèle une bande de forte intensité correspondant à un fragment d'environ 1Mb. S'agit-il du chromosome entier migrant plus haut qu'attendu, d'un marquage aspécifique ou bien d'une limitation de la résolution du gel? Cela pourra être confirmé ultérieurement par l'hybridation d'autres sondes, spécifiques du gène *ND7* ou du « bras gauche » du chromosome. Il est aussi possible de changer les conditions de migrations pour être plus résolutifs dans les haut poids moléculaires.

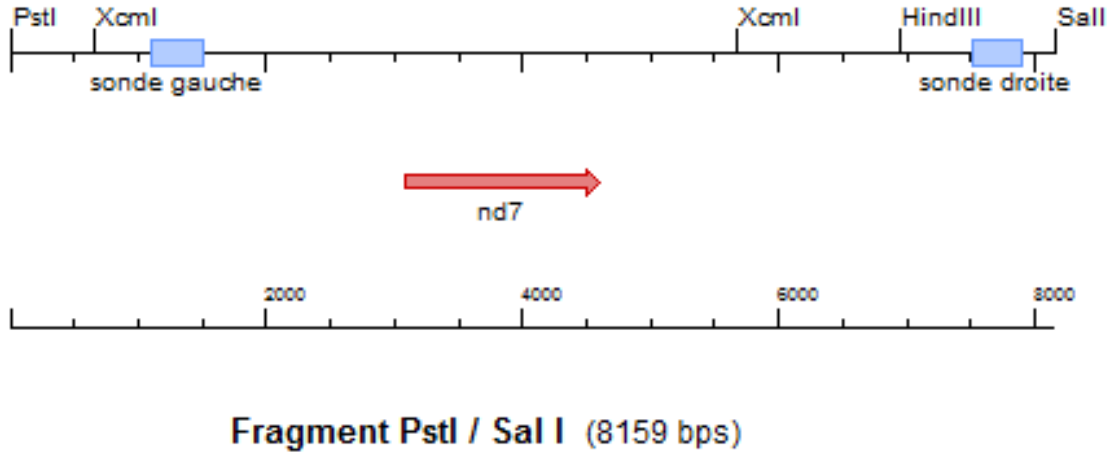


Schéma du gène *ND7* dans le fragment de restriction PstI/SalI et localisation de la sonde droite spécifique du « bras droit » du scaffold 5.

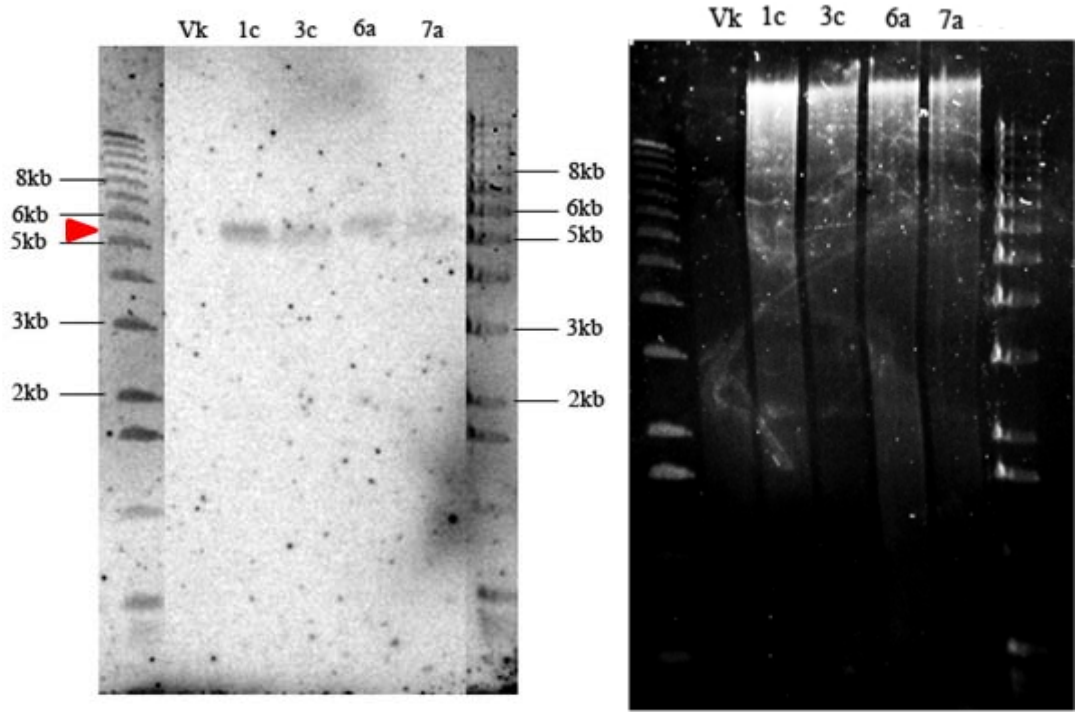


Figure 45 : Southern Blot (à gauche) et coloration BET (à droite) de l'électrophorèse en condition classique de fragment de restriction PstI/SalI sur lequel a été hybridée une sonde spécifique du bras droit, marquée radioactivement.

A l'échelle du gène

En parallèle les mêmes ADN ont été digérés par les enzymes PstI et SalI et mis à migrer sur gel d'agarose puis transférés sur membrane sur laquelle a été passée une sonde marquée spécifique du coté droit, par rapport au gène *ND7* (figure 45). On observe une bande diffuse entre 5 et 6kb qui pourrait correspondre au fragment PstI/SalI délété de *ND7* tandis que dans la piste correspondant à l'ADN d'un clone sauvage en croissance végétative on n'observe pas cette bande. On s'attendait dans cette piste à détecter une bande discrète de 8kb, ce qui n'est pas le cas. Toutefois, un marquage BET révèle que dans cette piste, la quantité d'ADN déposée était très faible, l'absence de toute bande n'est donc pas surprenante. Il faudra refaire cette expérience pour confirmer les résultats.

Addition des télomères aux extrémités

Pour tester l'addition des télomères au niveau des extrémités de chaque coté de la délétion induite du gène *ND7* j'ai réalisé des PCR dites « télomériques » (figure 46). J'ai utilisé des oligonucléotides spécifiques de l'ADN MAC flanquant de chaque coté le gène ND7 et un oligonucléotide s'appariant avec les séquences télomériques. Par cette technique il est possible de détecter un produit sous forme de smear après marquage BET. Un transfert sur membrane et l'utilisation d'une sonde oligonucléotide en aval des oligonucléotides utilisés pour la PCR confirme qu'il s'agit bien d'une amplification spécifique.

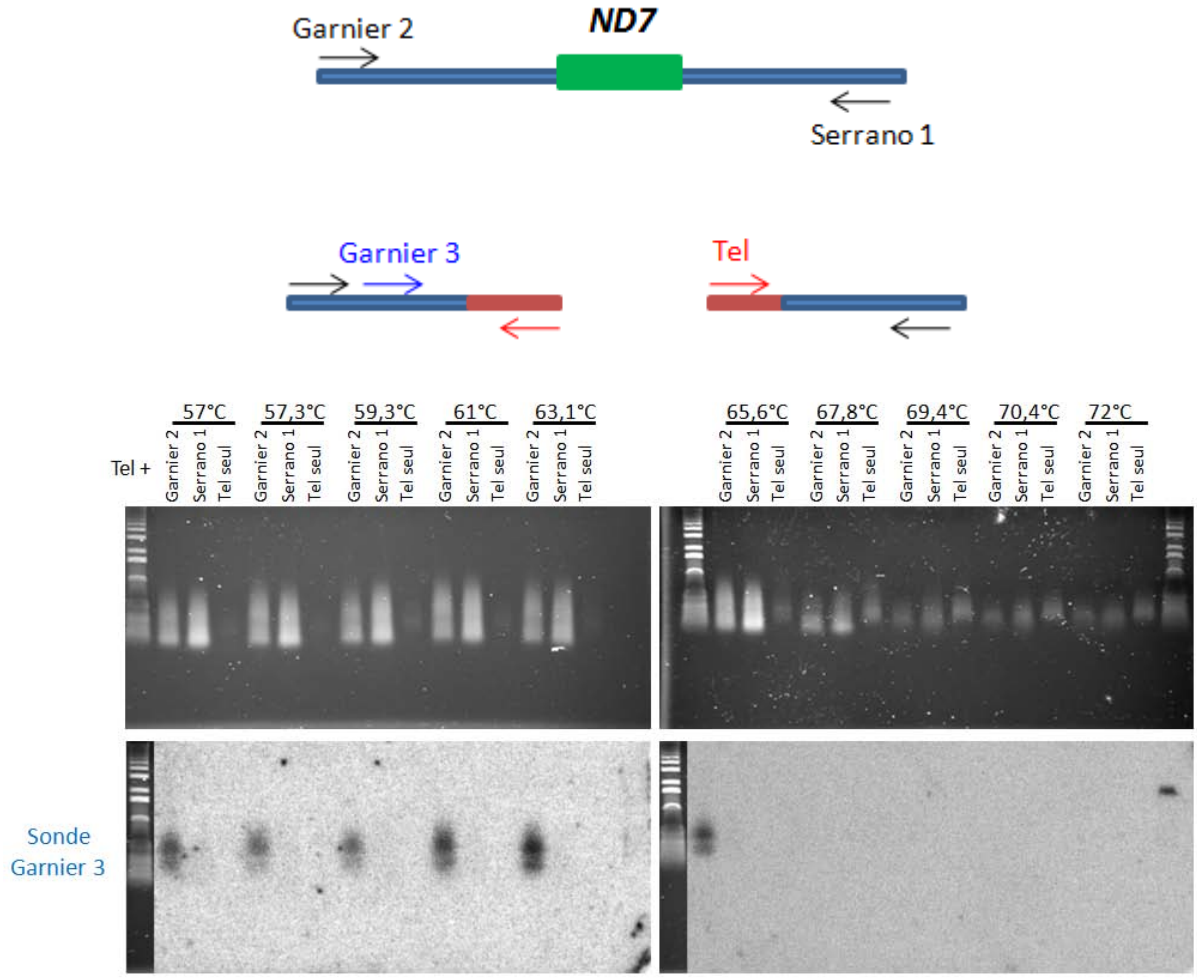


Figure 46 : PCR télomériques de chaque coté de la délétion induite du gène *ND7*. En haut est indiqué la localisation des différents oligonucléotides utilisés. En bas sont affichés les résultats marqués au BET, et le Southern Blot avec la sonde « Garnier 3 » marquée radioactivement.

Conclusions

Les méthodes complémentaires d'électrophorèse classique et d'électrophorèse en champ pulsé ont permis de caractériser le site de la délétion induite et héritable du gène ND7. Bien qu'il soit possible d'obtenir des résultats plus précis en affinant les conditions des expériences, on a pu confirmer que cette délétion induite se comporte comme une région naturelle de réarrangements imprécis. Il est possible d'y observer des délétions internes de taille variables, de la fragmentation des chromosomes et l'addition de télomères aux extrémités.

DISCUSSION

LES IES

Les IES dérivent de transposons

Contraintes topologiques lors de l'excision des IES

Quelles marques épigénétiques pour l'excision des IES

Différents Pgm-like pour différentes IES ?

MECANISME D'EXCISION DES IES

Un mécanisme de couper-coller

Complexe de pré excision Ku-Pgm

Spécialisation des protéines Ku

Maturation des extrémités d'ADN dépendante du complexe de ligation

VERS UN PROCESSUS INTEGRE

Une nouvelle vision de la voie NHEJ classique

L'excision des IES est un processus hautement intégré

Comparaison avec *Tetrahymena thermophila*

Pourquoi un mécanisme différent chez *Tetrahymena thermophila*?

DOMESTICATION DES REARRANGEMENTS PROGRAMMES

Boucle de contrôle des réarrangements programmés

Un candidat pour le contrôle des réarrangements programmés

LE NHEJ ALTERNATIF POUR L'ELIMINATION HETEROGENE

LES IES

Les IES dérivent de transposons.

L'analyse du jeu de 45 000 IES identifiées grâce au séquençage du génome de MAC en développement de cellules inactivées pour PGM a montré qu'au moins une partie des IES dérivent de transposons Tc1/mariner, en accord avec l'observation d'un consensus dégénéré aux bornes des IES et des éléments Sardine, Thon et Anchois trouvés dans le génome germinal de *Paramecium tetraurelia*. En effet, certaines IES parmi les plus longues identifiées, typiquement d'une taille supérieure à 900 pb correspondent à des séquences inversées répétées et pourraient être d'anciens pieds de transposons dégénérés (Arnaiz et al., 2012).

L'analyse des duplications globales du génome qui ont eu lieu au cours de l'évolution de *Paramecium tetraurelia* montre que les IES sont apparues progressivement et qu'elles tendent à être plus courtes lorsqu'elles sont apparues plus tôt. Cela suggère plusieurs vagues d'invasion du génome par des éléments mobiles.

Pgm est une transposase domestiquée de type *piggybac*, cependant l'analyse des 45000 IES du jeu de référence montre que la séquence TTAA coupée par la transposase Piggybac est clairement sous-représentée aux bornes des IES. Seuls la géométrie des cassures introduites par Pgm et sa spécificité pour le dinucléotide TA central commun aux éléments Tc1/mariner et piggybac ont été conservés pendant l'évolution. Cependant si on regarde la distribution des bornes selon les tailles des IES, on observe des différences considérables. Par exemple les bornes ATAC sont significativement sur représentées pour les IES d'une taille supérieure à

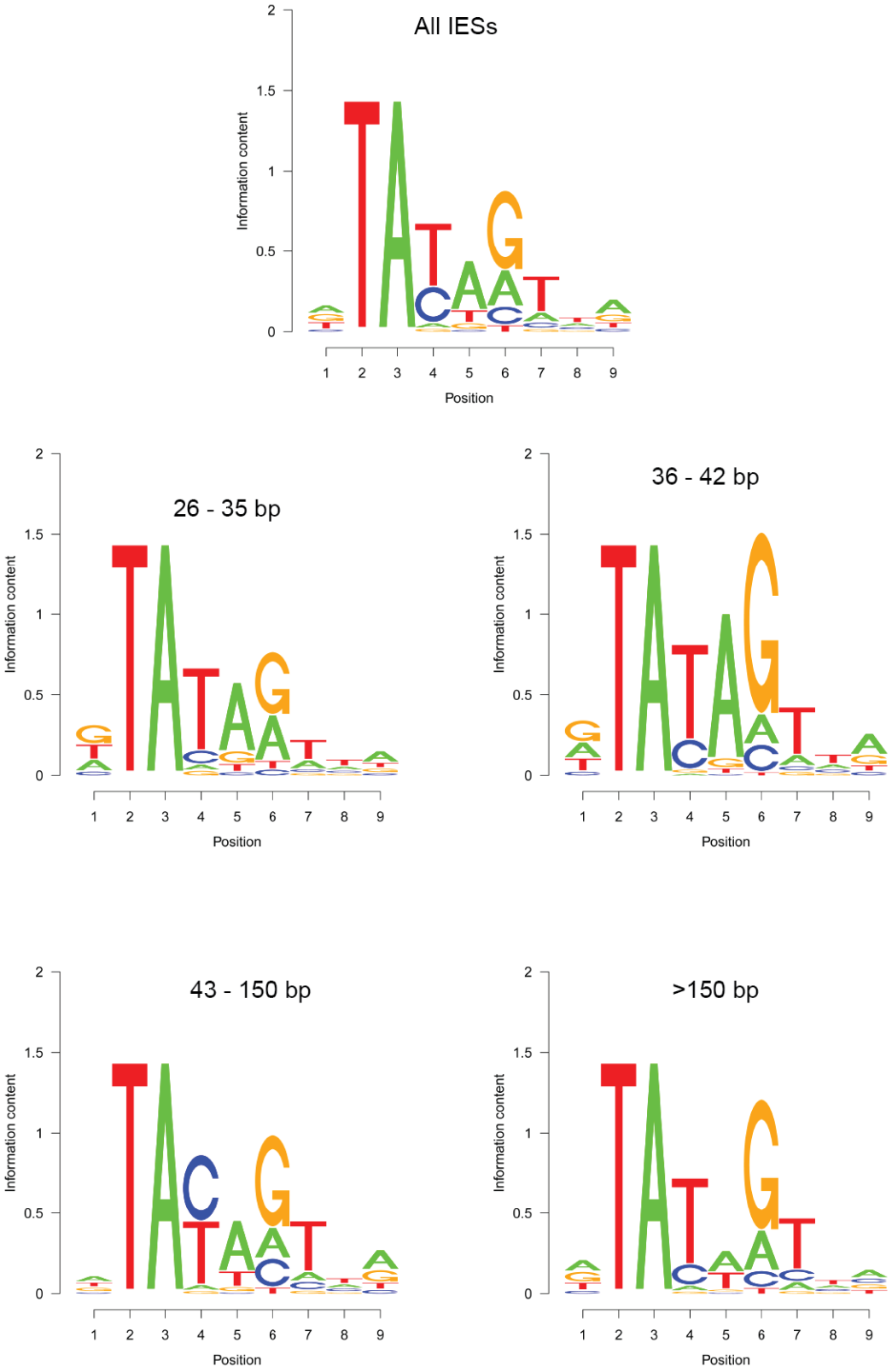


Figure 47 : Logo des séquences observées aux bornes des IES, un nucléotide en amont du TA central et six nucléotides en aval. Les résultats sont présentés pour l'ensemble des IES ou classés par catégories de taille (Linda Sperling, communication personnelle).

35pb, particulièrement celles qui ont une taille comprise entre 43 et 150pb. Dans cette même catégorie de taille, la borne GTAT, qui représente toujours une grande majorité des bornes d'IES sur-représentées par rapport aux fréquences théoriques, est ici moins favorisée. Cette observation est cohérente avec les analyses faites dans d'autres laboratoires. Les analyses s'intéressaient au consensus de séquences retrouvés à l'intérieur des IES à partir du dinucléotide TA (figure 47). Il a été observé que les IES de taille intermédiaires, entre 43 et 150 pb, avait un biais au niveau du nucléotide situé après le TA. Dans cette catégorie de taille, le T est moins favorisé tandis que le C est plus présent (communications personnelles de Linda Sperling et Estienne Swart). Une analyse extensive des séquences des bornes d'IES, peut-être sur une plus grande distance, en fonction de caractéristiques telle que la taille, la dépendance à certains facteurs d'excision, fournira sans aucun doute d'intéressantes informations sur l'évolution des IES et les différentes catégories d'IES excisées.

Contrainte topologique lors de l'excision des IES

Une grande majorité des IES ont une taille qui tend vers 26-32 pb et le dinucléotide TA semble être la seule signature qui soit absolument conservée aux bornes de toutes les IES. L'analyse de la distribution en taille des IES montre une périodicité d'environ 10 pb, soit un tour d'hélice d'ADN. Cette périodicité tend à disparaitre lorsque les IES sont de plus grande taille.

Une autre observation intéressante est le fait qu'il y a une rupture dans la distribution de taille des IES. Les IES d'une taille comprise entre 33 et 43pb, qui pourtant pourraient correspondre à un pic de périodicité sont clairement sous représentées

Il a été montré que l'introduction des cassures double brin aux bornes des IES nécessitait un dialogue entre les deux bornes. Une mutation dans le dinucléotide TA d'un coté de l'IES

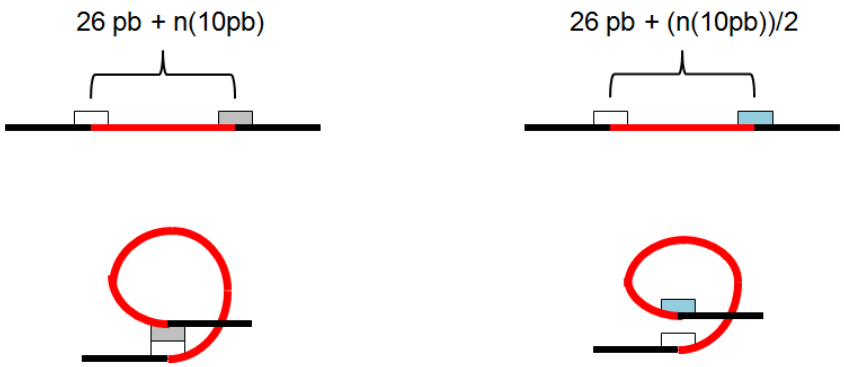


Figure 48 : Contrainte de phase dans le mécanisme d'excision des IES. Les IES (ici en rouge) en phase sont représentées à gauche, hors phase à droite. Les bornes des IES sont représentées par les rectangles blancs et gris.

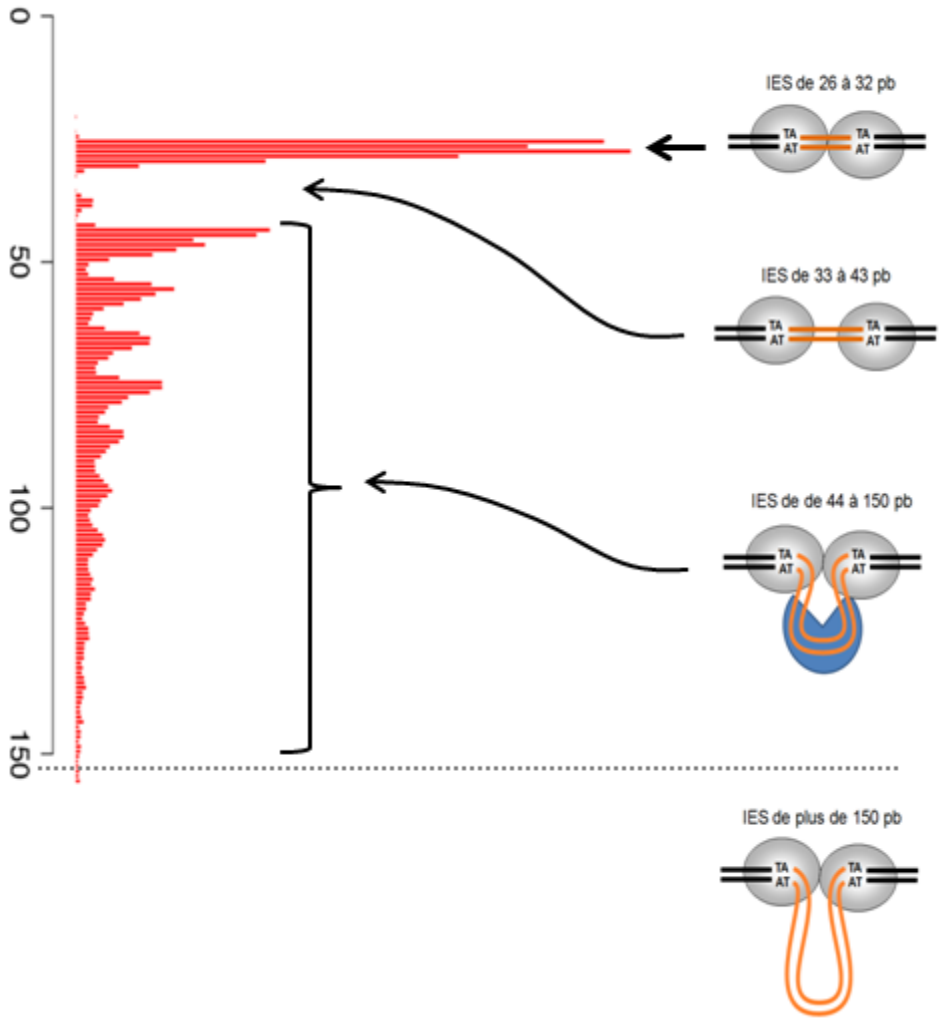


Figure 49 : Contrainte sur le mécanisme d'excision des IES. Modèle pour l'utilisation d'une protéine capable de courber l'ADN pour les IES d'une taille comprise entre 44 et 150 pb.

prévient l'introduction des cassures des deux cotés. Les transposases ayant tendance à agir en multimère, on peut imaginer que les protéines Pgm qui introduisent les cassures aux bornes des IES doivent interagir pour être actives.

Toutes ces observations orientent vers un modèle d'excision des IES mettant en jeu un rapprochement physique des deux extrémités. Ce modèle induirait des fortes contraintes topologiques pour la formation de la synapse permettant le rapprochement des extrémités d'IES.

Le mécanisme d'excision des IES est peut-être plus efficace pour les IES dont la taille correspond à la périodicité observée sur l'histogramme de distribution, les IES se trouvant dans les creux, avec un décalage d'un demi-tour d'hélice pourraient ne pas s'agencer correctement en phase dans le mécanisme d'excision (figure 48).

La sous représentation de la classe d'IES d'une taille comprise entre 33 et 43pb suggère l'intervention d'une protéine capable de courber l'ADN dans le modèle de rapprochement physique des extrémités via la formation d'une synapse (figure 49).

Dans ce modèle, les IES ayant une taille autour de 26-32 pb seraient suffisamment courtes pour former un complexe d'introduction des cassures sans nécessité de courber l'ADN, les protéines Pgm impliquées dans le complexe étant suffisamment grosses pour interagir directement.

Les IES d'une taille comprise entre 44 et 150pb pourraient former la synapse permettant le dialogue entre les deux bornes grâce à l'action d'une protéine permettant la courbure de l'ADN. Les IES en décalage de phase seraient également capables de rapprocher les bornes d'IES mais seraient incapables de former un complexe d'introduction de cassures stable et efficace, et auraient été contre-sélectionnées au cours de l'évolution. Peut-être le biais de

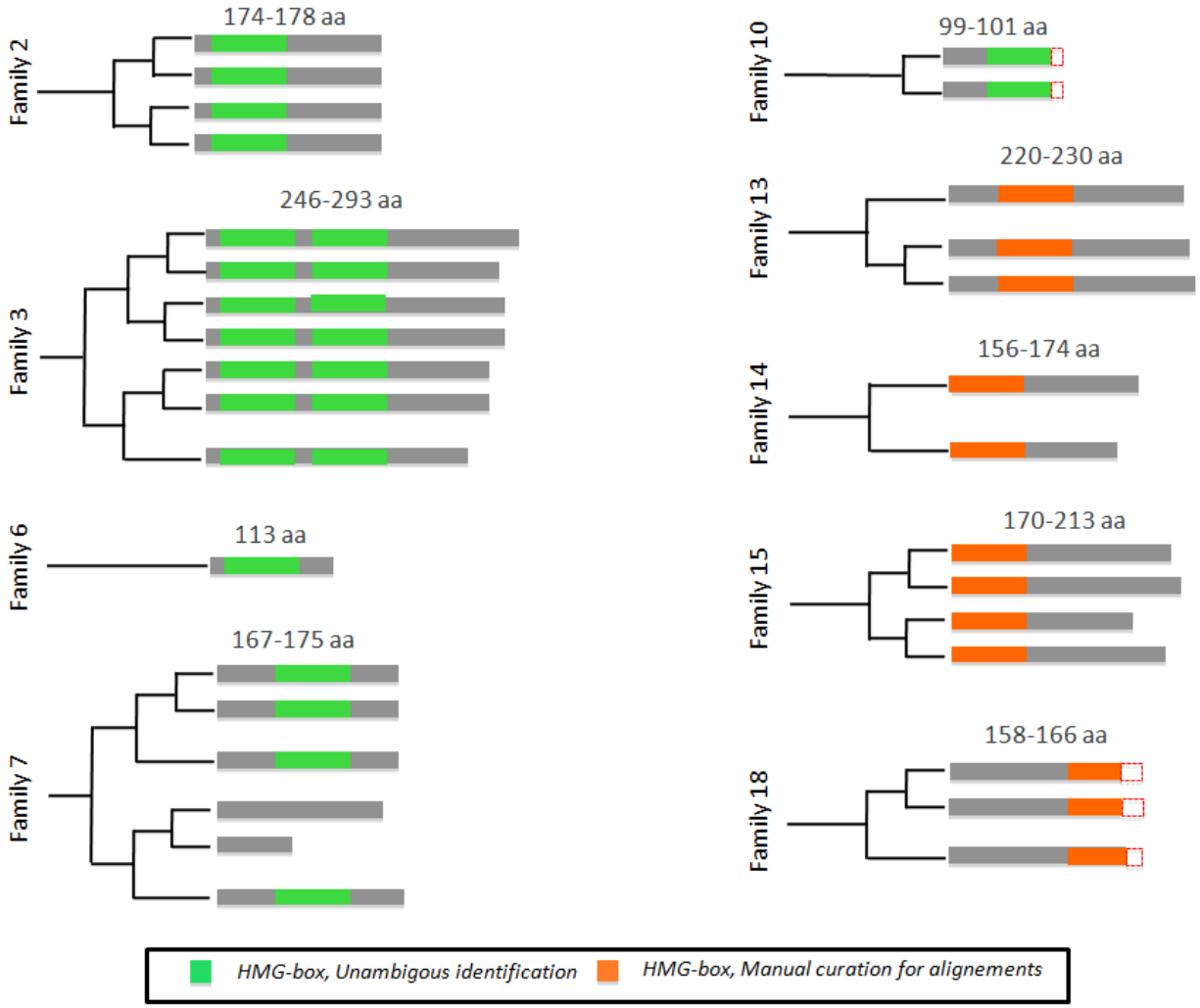


Figure 50 : Gènes codant pour des protéines contenant des domaines HMGB et significativement induits pendant les réarrangements programmés du génome. Les familles indiquent les relations d'ohnologie. En vert sont les domaines HMGB annotés automatiquement. En orange ceux qui ont nécessité une annotation manuelle (Vinciane Regnier, communication personnelle).

séquence observé aux bornes de ces IES est lié à cette contrainte topologique ; cette séquence pourrait favoriser la courbure de la molécule d'ADN.

Pour les IES d'une taille comprise entre 33 et 43 pb, on imagine que ces IES sont, d'une part trop longues pour permettre une interaction directe entre les complexes de coupure de chaque coté de l'IES, et d'autre part trop courtes pour pouvoir être courbées et permettre le dialogue entre les deux bornes, même avec le concours d'une protéine capable de courber l'ADN.

Il existe chez les eucaryotes une famille de protéines connue pour se lier à la molécule d'ADN et la courber. Ces protéines HMG (pour High Mobility Group) sont impliquées dans de nombreux processus cellulaires. Plus particulièrement les protéines HMGB, sont impliquées dans la recombinaison V(D)J et il a été montré qu'elles étaient importantes pour la formation de la synapse 12/23 RSS et le clivage par les protéines RAG (Ciubotaru et al., 2013).

Cette famille de protéines est présente chez la paramécie (figure 50). 53 gènes répartis en 18 familles, codant pour des protéines de type HMGB ont été identifiés, certains sont spécifiquement induits pendant les réarrangements programmés du génome. L'analyse par spectrométrie de masse des protéines présentes dans des préparations de noyaux enrichis en macronoyaux en développement montre qu'au moins une partie de ces protéines est présente. Il a été montré que les gènes de la famille 2 ont un rôle dans la progression de la méiose, lorsqu'ils sont inactivés par ARN interférence, il n'y a pas de développement de nouveaux MAC. L'unique gène de la famille 6 n'est pas essentiel, son inactivation pendant l'autogamie n'a aucun effet notable sur la descendance. Enfin l'extinction d'un gène de la famille 13 montre un léger effet avec 30% de mortalité dans la descendance post autogame. (Vinciane Regnier, communication personnelle)

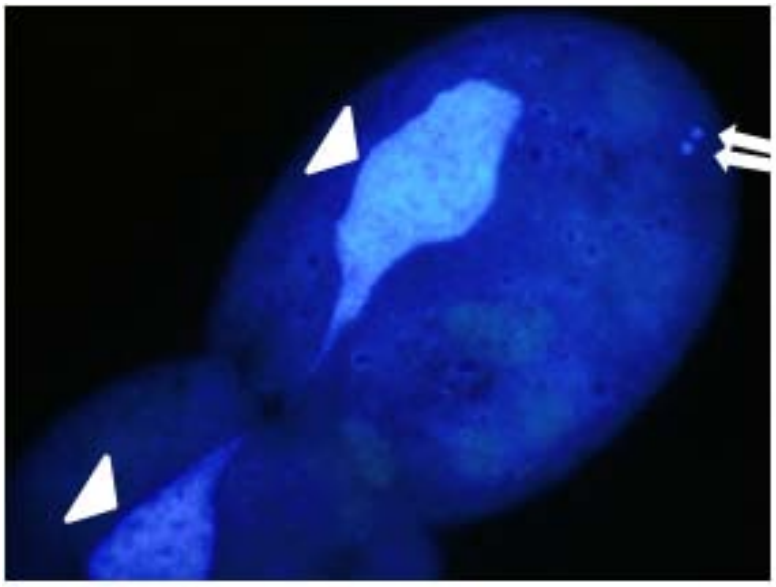


Figure 51 : Une division amitotique de paramécie dont les noyaux sont marqués au DAPI. Les triangles indiquent les macronoyaux, les flèches indiquent les micronoyaux.

Quelles marques épigénétiques pour l'excision des IES ?

Les IES dérivent de transposons, au moins pour une partie d'entre elles, mais pas forcément de transposons PiggyBac. Aucun motif n'est absolument conservé au bornes des IES en dehors du dinucléotide TA central. Le génome de la paramécie est naturellement riche en TA, environ 70% pour les séquences MAC, jusqu'à 80% pour les séquences d'IES, le ciblage des séquences à éliminer n'est donc pas uniquement dépendant de ce TA. Ce ciblage doit aussi être dépendant de marques « épigénétiques ».

On peut imaginer que ces marques soient des variants d'histones spécifiques de certaines séquences, ou encore que ces histones portent des modifications particulières (phosphorylation, acetylation, méthylation et ubiquitination) Dans le génome de paramécie, il existe de nombreux gènes codant pour des protéines capables de modifier les histones.
Des gènes codant pour des histones méthyltransférases de type Ezl sont présentes, l'un d'entre eux est spécifiquement induit pendant la méiose et est nécessaire pour l'élimination d'une partie des IES (Sandra Duharcourt, communication personnelle).

La division de la paramécie est particulière. Si les mics se divisent par mitose classique, le macronoyau quant à lui se divise de façon amitotique (figure 51). On peut légitimement se demander ce qui différencie ces noyaux et qui pourrait expliquer les différences lors de la division. Une hypothèse évidente est que le macronoyau n'a pas les centromères indispensables pour une division classique, et que ces centromères micronucléaires pourraient être éliminés physiquement pendant les réarrangements programmés du génome..
Tous les organismes possèdent des variants d'histones spécifiques des centromères. Et l'équipe de Sandra Duharcourt est parvenu à identifier le gène codant pour CenH3, le variant

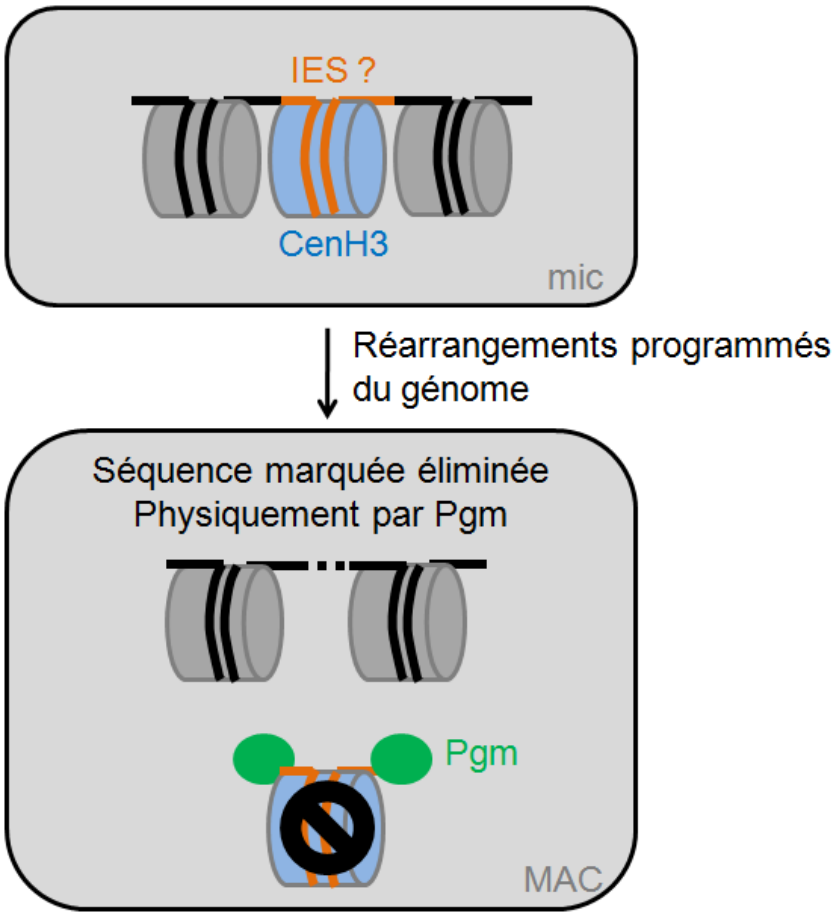


Figure 52 : Modèle d'élimination des centromères pendant les réarrangements programmés du génome par la protéine Pgm. La nature des séquences centromériques éliminées est encore inconnue.

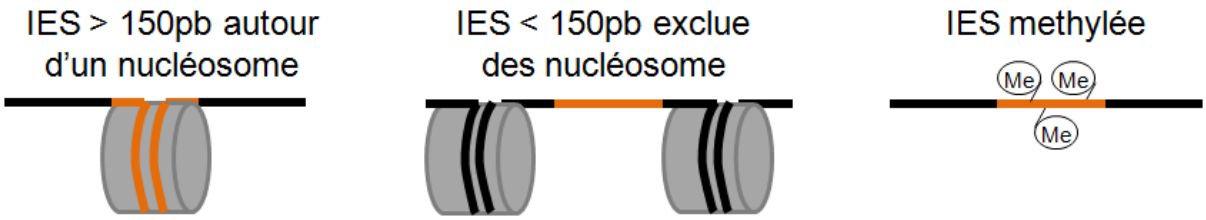


Figure 53 : Modèles pour les différentes façons de cibler les séquences à éliminer par la machinerie d'excision. Les séquences éliminées apparaissent en orange, les séquences flanquantes en noir.

centromérique de l'histone H3 de paramécie. Des anticorps reconnaissant CenH3 marquent spécifiquement les micronoyaux. Lorsque la protéine Pgm est déplétée, les anticorps marquent également les macronoyaux en développement (Sandra Duharcourt, communication personnelle). Cela suggère que les séquences centromériques, portant les variants cenH3 sont reconnus par la machinerie d'élimination et sont physiquement éliminées des macronoyaux lors des réarrangements programmés du génome (figure 52). Les séquences « centromères » n'ont pas encore été identifiées mais on peut imaginer qu'il s'agit d'IES ou d'anciens éléments mobiles.

Mais qu'en est-il de la grande majorité des IES dont la taille est inférieure à 150 pb? Il semble impossible de cibler les séquences de petites tailles via des variants ou des marques d'histones spécifiques. Deux hypothèses peuvent être imaginées :

Une possibilité serait que les IES de petite taille soient exclues des nucléosomes et ce serait l'absence d'histones justement qui ciblerait les séquences à éliminer. On peut aussi imaginer des marques spécifiques pour les histones qui encadrent les IES pour diriger le complexe d'excision vers ces séquences exclues des nucléosomes (figure 53). Une analyse alliant approche bioinformatique et séquençage haut débit des séquences associées aux nucléosomes de la paramécie au laboratoire (collaboration avec Sandra Duharcourt et Linda Sperling) est en cours et pourrait nous en apprendre plus sur le positionnement des nucléosomes par rapport aux séquences à éliminer.

Une autre possibilité pourrait être la méthylation de la molécule d'ADN elle-même. La méthylation joue un rôle dans divers processus cellulaires : la synchronisation de la réplication du chromosome chez les bactéries, la réparation des mésappariements dans l'ADN et aussi sur le niveau d'expression de certains gènes. Chez le cilié *Oxytricha trifallax*, la méthylation de novo des cytosines est observée uniquement pendant la conjugaison, pendant

les réarrangements programmés du génome. On trouve ces modifications sur des séquences de transposons, des répétitions minisatellites, et d'autres séquences spécifiques de la lignée germinale. L'utilisation de drogues inhibant la méthylation des cytosines résulte en la rétention des séquences à éliminer dans les cellules après conjugaison (Bracht et al., 2012).

Chez *Tetrahymena* et *Paramecium* cependant on ne détecte pas de méthylation des cytosines (Karrer and VanNuland, 2002). L'analyse qui a été mené chez *Tetrahymena* indique qu'il existe des sites des méthylation des adénines et qu'il semble que cette méthylation se situe principalement dans les régions « linker », donc exclues des nucléosomes.

Moins d'une dizaine de gènes pouvant coder pour des adénines méthytransférases a été trouvé par blast sommaire dans le génome de *Paramecium tetraurelia*. Néanmoins aucun d'entre eux n'a de profil d'expression suggérant une implication dans les réarrangements programmés du génome.

Différents Pgm-like pour différentes IES ?

Il reste encore beaucoup à découvrir pour expliquer les différences observées lors de l'excision des IES et établir une classification des 45000 IES identifiées en fonction de leurs exigences pour être excisées. En effet, l'étude de la cinétique de coupure aux bornes et les analyses par PCR montrent que certaines IES sont excisées plus précocement que d'autres, certaines sont sous contrôle maternel fort et d'autres non (Duharcourt et al., 1998; Nowacki et al., 2005). L'équipe de Linda Sperling au Centre de Génétique Moléculaire utilise les données de séquençage haut débit pour identifier les IES sensibles à l'extinction de différents facteurs impliqués dans le contrôle de l'excision, Nowa (Nowacki et al., 2005), Ptiwi (Bouhouche et al., 2011) par exemple, et vérifier s'il y a une corrélation avec les bornes observées, la taille, etc… L'ADN extraits aux temps tardifs de développement MAC à partir de cellules déplétées pour ces facteurs a été analysé et on a découvert que certaines IES étaient plus affectées

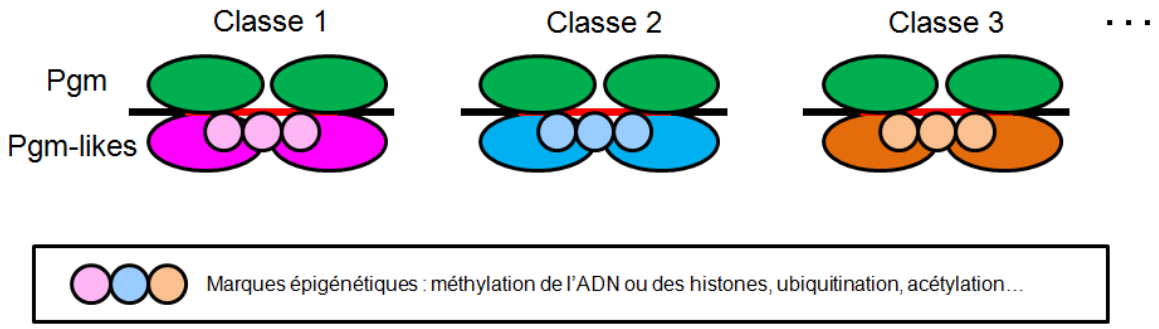


Figure 54 : Modèle pour un complexe de cassure Pgm/Pgmlike. Différents Pgm-likes reconnaissent les différentes marques épigénétiques qui définissent les classes d'IES.

par ces extinctions, permettant une ébauche de classification des IES (Cyril Denby-Wilkes, communication personnelle). Cela suggère que cohabitent dans les MAC en développement plusieurs mécanismes d'excision d'IES ayant pour cœur Pgm et la voie NHEJ. En effet toutes les IES dépendent de Pgm, Ku ainsi que de la ligase IV et de XRCC4 pour l'excision. La présence de nombreux gènes *PiggyMac-like* dans le génome de *Paramecium tetraurelia* pourrait contribuer à la variabilité entre IES. En effet, certains de ces gènes, bien que codant pour des protéines ne possédant pas la triade catalytique DDD caractéristique sont indispensables aux cellules pour donner une descendance post autogame viable (résultats non publiés, labos Bétermier et Nowacki). Des analyses moléculaires en cours montrent que les protéines correspondantes sont nécessaires pour l'excision des IES. Les transposases ayant tendance à agir sous forme de multimères, on peut imaginer que l'association entre Pgm et un Pgm-like sont spécifiques d'une catégorie d'IES. Si ces protéines ne possèdent pas la triade catalytique, peut-être leurs domaines additionnels permettent la reconnaissance de marques épigénétiques ou de certains motifs de séquence caractéristiques de catégories particulières d'IES, qui aurait échappé à l'analyse globale.

Des expériences préliminaires ont été menées par Julien Bischerour au laboratoire pour tester les marques épigénétiques reconnues in vitro par la protéine Pgm et par certaines Pgm-likes. Effectivement Pgm et Pgm-like4 reconnaissent des modifications sur les histones, certaines étant différentes entre les deux protéines. Ainsi on peut construire un modèle où une catégorie d'IES, signalée par un assortiment particulier de marques épigénétiques, serait reconnue par un oligomère composé de Pgm, pour l'activité catalytique, et d'un Pgm-like, pour la reconnaissance des marques spécifiques (figure 54). Le séquençage haut débit de l'ADN de cellules déplétées pour les différents Pgm-like permettra d'examiner à l'échelle du génome si chaque Pgm-like intervient dans l'excision d'une classe particulière d'IES ; cela permettra

aussi de rechercher des déterminants de séquences particuliers présents dans cette classe.

Il serait également informatif d'utiliser des anticorps reconnaissant les modifications d'histones reconnues in vitro par Pgm et les Pgm-likes pour des expériences de ChIP Seq, afin de vérifier l'association des IES de chaque classe avec les modifications des histones.

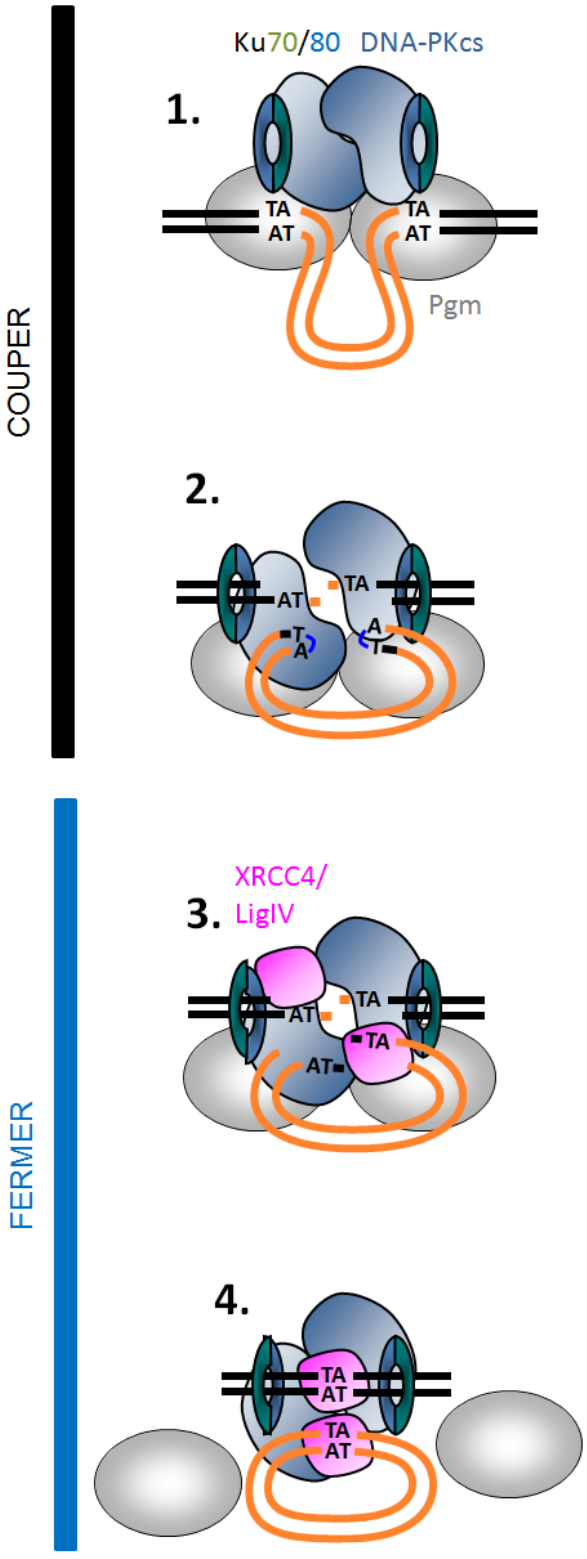


Figure 55 : Mécanisme de couper-fermer avec un complexe Pgm-DNA-PK. Dans ce modèle l'introduction des cassures est assurée par un complexe Pgm-DNA-PK, et une refermeture de la lésion par la voie NHEJ classique.

Mécanisme d'excision des IES

Un mécanisme de « couper-fermer »

L'ensemble des résultats sur l'excision des IES et la voie NHEJ indique que l'excision est un mécanisme de « couper-fermer ». En effet le processus d'excision peut être divisé en deux étapes. L'introduction de cassures double brin des deux cotés des IES de façon concertée, puis la refermeture précise de la cassure par la voie NHEJ (figure 55). En cela il peut être comparé à la transposition de type « couper-coller » à ceci près que chez la paramécie les IES et les transposons sont activement et physiquement éliminés pendant la formation du nouveau MAC et ne sont pas réinsérés dans le génome.

Complexe de pré excision Ku-PGM

Les résultats montrent que lorsque l'hétérodimère Ku est absent lors de l'excision des IES, il n'y a pas introduction des cassures par Pgm aux bornes des séquences à éliminer. On a montré que les protéines Ku80c et Pgm, spécifiquement induites pendant les réarrangements programmés du génome interagissent, fonctionnellement car l'extinction de KU80c affecte la localisation de Pgm, et physiquement comme cela a été montré par les expériences de co-purification et de colocalisation.

Lors des réarrangements programmés du génome, la paramécie doit exciser plus de 45 000 IES dont la moitié interrompt des cadres ouverts de lecture. La cellule doit donc dans un laps de temps relativement court, gérer l'introduction de plus de 90 000 cassures double brin et

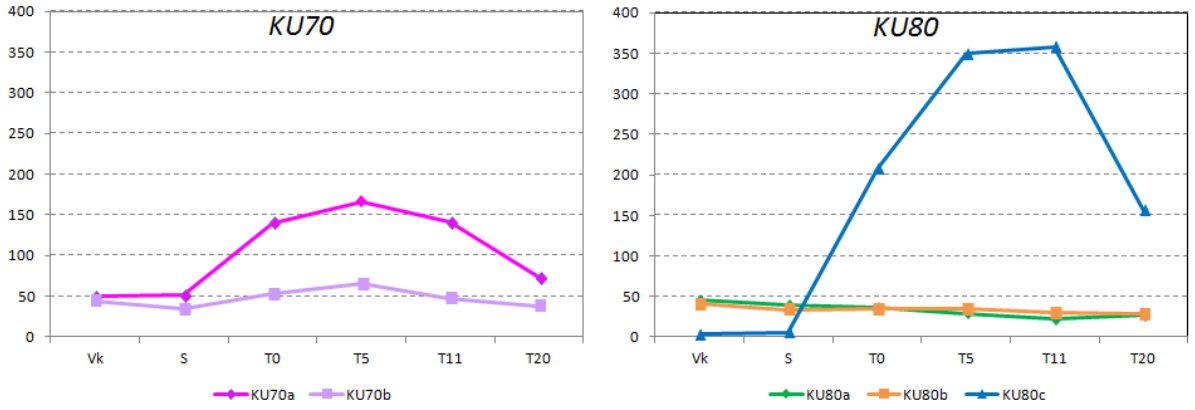


Figure 56 : Profil d'expression des gènes Ku lors des réarrangements programmés du génome. Données de séquençage des ARN messagers lors d'une cinétique d'autogamie.

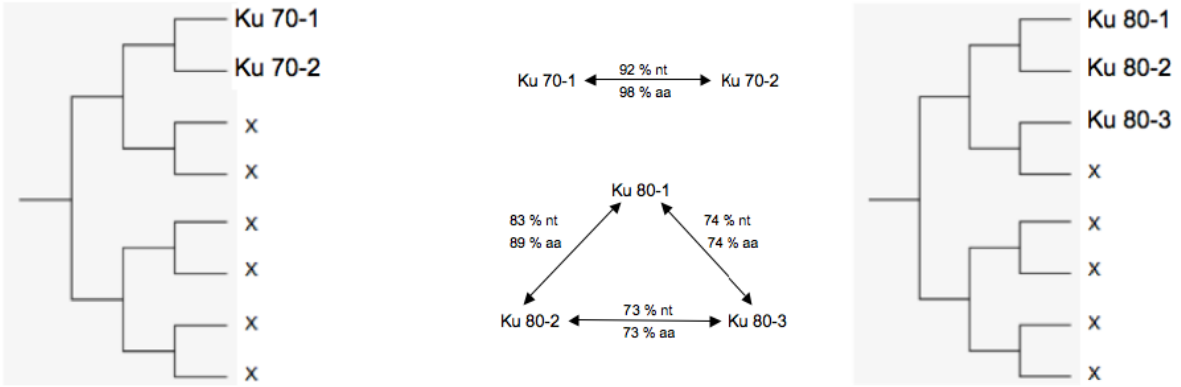


Figure 57 : Relation d'ohnologie entre les différents gènes Ku et pourcentages d'identité. Les X indiquent les paralogues perdus lors des résolutions des trois duplications globales successives du génome.

45 000 événements de réparation par le NHEJ, précis au nucléotide près, sans tenir compte du niveau de ploïdie atteint dans le MAC en développement au moment où débutent les réarrangements. Il s'agit ici d'une différence fondamentale avec la recombinaison V(D)J. Les réarrangements programmés du génome chez la paramécie ont lieu à l'échelle de tout le génome et doivent être extrêmement précis. Il n'est donc pas illogique de penser que pour faire face à ce défi moléculaire, éviter les translocations de chromosomes et s'assurer que les extrémités ADN ne seront pas la cible de nucléases processives, la cellule ait sélectionné au cours de l'évolution un mécanisme qui couple étroitement introduction des cassures double brin et réparation.

Spécialisation des protéines Ku

Il y a trois gènes *KU80* et deux gènes *KU70* dans le génome de la paramécie. L'étude a montré que *KU80c* était absolument indispensable pour les réarrangements programmés du génome tandis que l'extinction de *KU80a* et *KU80b* n'a pas d'effet sur la descendance. Le gène *KU80c* est spécifiquement transcrit (figure 56) pendant le développement du macronoyau, la protéine encodée localise dans les macronoyaux en développement et interagit avec la protéine Pgm. Tous ces indices suggèrent une spécialisation fonctionnelle de la protéine Ku80c. On ne sait pas encore si cette spécialisation se fait uniquement par l'élévation du niveau d'expression pendant les réarrangements ou si la protéine Ku80c porte des fonctions spécifiques pour interagir et activer Pgm. Une expérience de sauvetage phénotypique, consistant à surexprimer Ku80a et/ou Ku80b, sous le contrôle des signaux de transcription de *KU80c*, pendant les réarrangements lorsque *KU80c* est éteint permettrait de répondre à cette question.

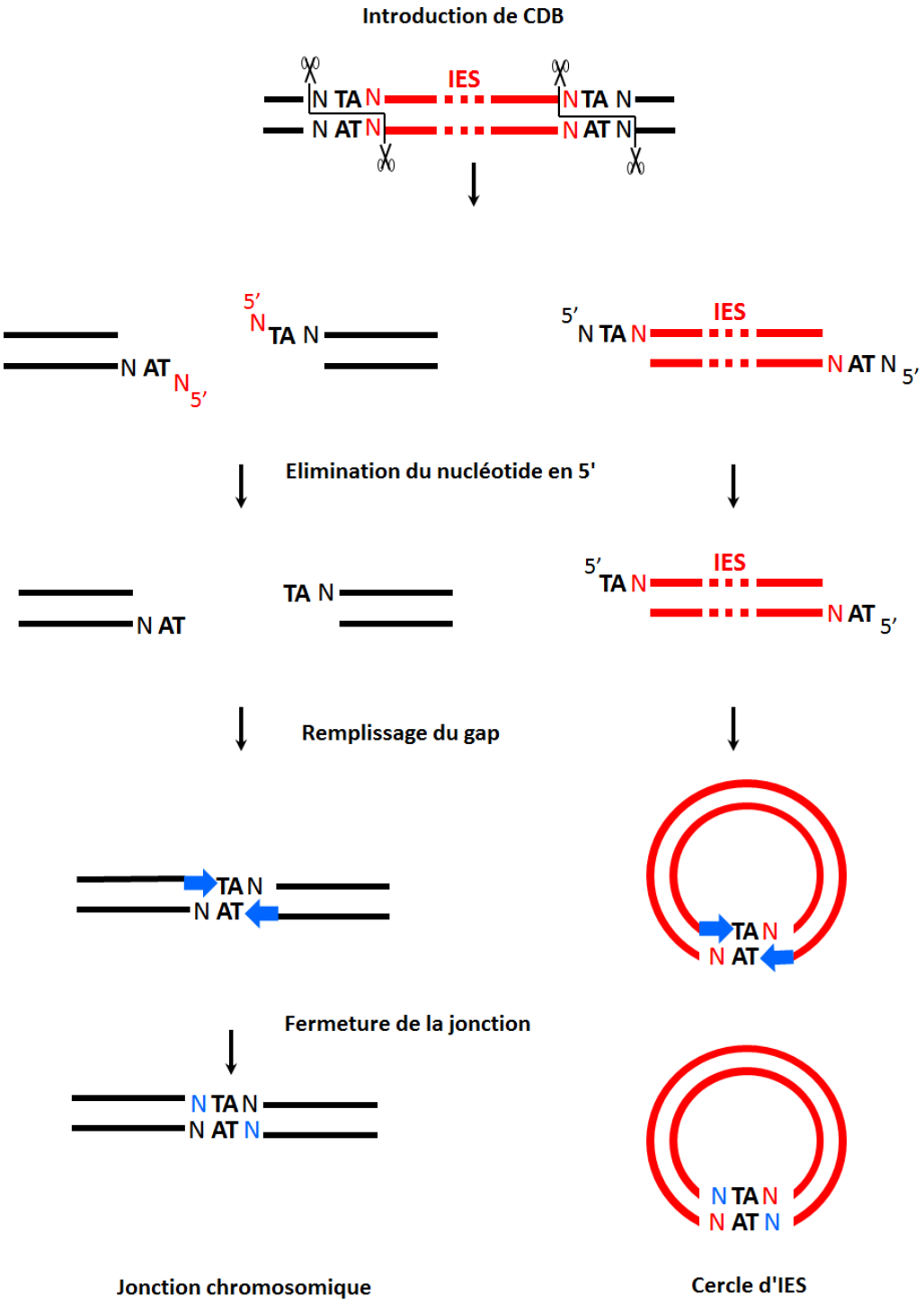


Figure 58 : Modèle d'excision des IES. Coté jonction chromosomique, et coté jonction d'IES. On imagine que les mécanismes sont les mêmes coté IES.

En ce qui concerne les gènes *KU70*, il est plus difficile de répondre à cette question. En effet ces deux gènes sont extrêmement proches et partagent 92% d'identité au niveau de leur séquence nucléotidique (figure 57). Ils ne peuvent donc pas être éteints séparément par ARN interférence. On sait qu'au moins un des gènes *KU70* est nécessaire. Cependant le gène *KU70a* est spécifiquement transcrit pendant les réarrangements programmés, et la protéine encodée par ce gène localise dans les macronoyaux en développement. Il est difficile d'imaginer que la protéine Ku70a porte des fonctions spécifiques via des résidus différents, peut-être la spécialisation entre ces deux gènes ne se traduit-elle ici que par une induction spécifique du gène *KU70a* pendant les réarrangements.

Maturation des extrémités dépendante du complexe de ligation

L'étude de l'implication des gènes *LIGIV* et *XRCC4* dans l'excision des IES (le modèle est rappelé sur la figure 58), en plus de prouver l'implication de la voie NHEJ dans les réarrangements programmés du génome a permis de mettre en évidence plusieurs points. L'introduction de cassures double brin programmées se fait aux deux bornes des IES. Les cassures programmées ne sont pas réparées en l'absence des protéines du complexe de ligation (Cernunnos n'a pas encore été testé mais est présent dans le génome de *Paramecium tetraurelia*) cependant ces extrémités ADN ne semblent pas dégradées, suggérant que l'hétérodimère Ku est fixé au niveau de ces cassures et les protège de la résection. Au vu des résultats sur l'étude de Ku, il est difficile de confirmer cette hypothèse, puisque Ku est nécessaire pour les cassures programmées dépendantes de Pgm, mais elle est cohérente avec le rôle supposé de l'hétérodimère Ku dans la maturation des extrémités lors de la réparation des cassures double brin. Ku est supposé avoir un rôle protecteur, empêchant la dégradation

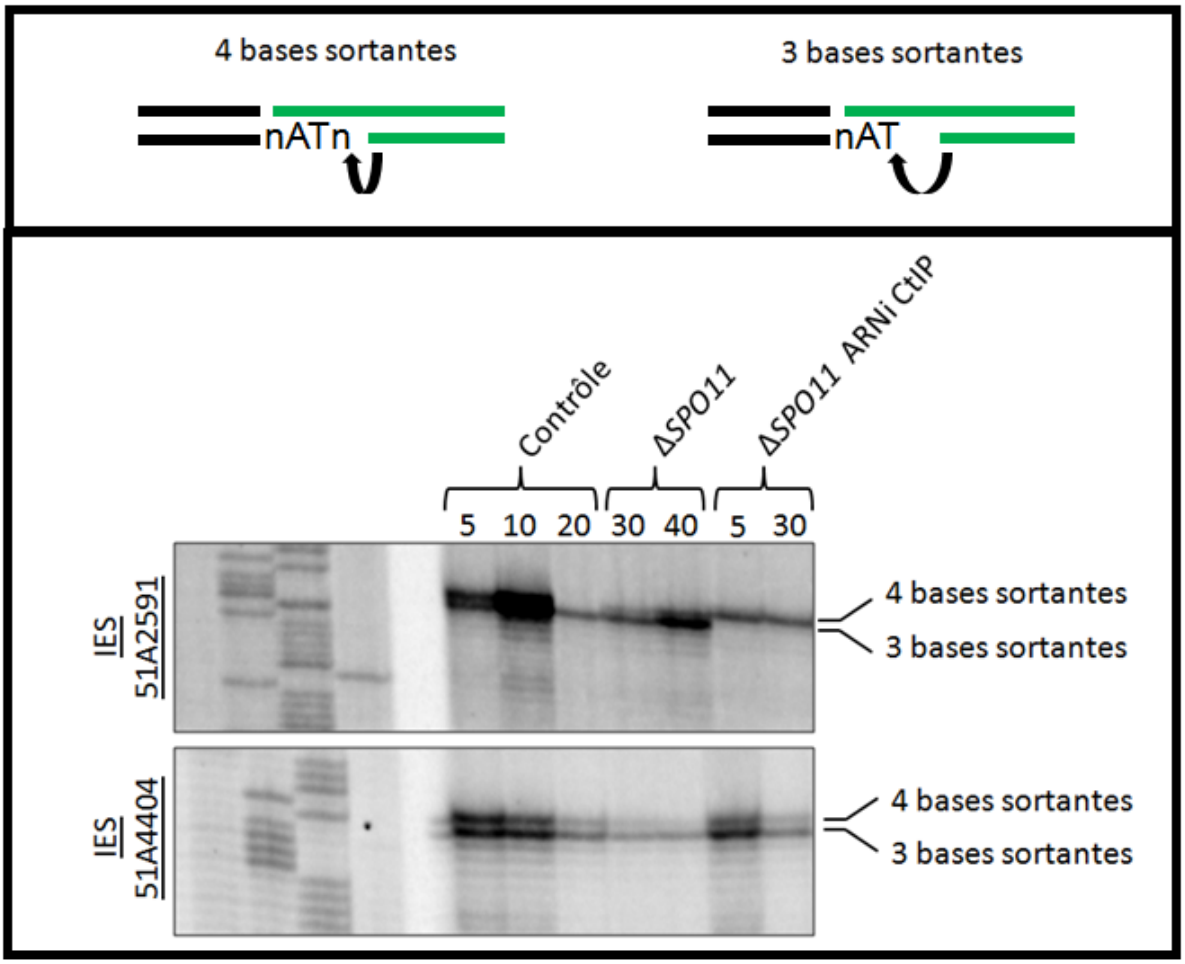


Figure 59 : Schéma des différents produits de ligation aux extrémités de 4 ou 3 bases sortantes en 5'. Maturation des cassures aux bornes des IES 51A2591 et 51A4404. La maturation de l'IES 5112591 semble dépendante de CtIP. Une analyse à plus grande échelle est nécessaire pour confirmer ces premiers résultats.

extensive des extrémités ADN par des nucléases processives comme Exo1 (Tomimatsu et al., 2012). Cependant, on sait que les extrémités 5' de quatre bases sortantes générées lors de l'introduction des cassures double brin par PiggyMac ne sont pas toujours complémentaires de part et d'autre du site d'excision. Il faut donc nécessairement des étapes de maturation des extrémités. Le modèle actuel propose une résection du dernier nucléotide en 5' pour permettre l'appariement des dinucléotides TA, suivi d'une étape de polymérisation de 5' vers 3' pour remplir le gap et permettre la religation de l'ADN.

Fort heureusement, la technique de LMPCR permet une ligation de l'adaptateur sur une extrémité de 4 bases sortantes, mais aussi sur une extrémité de 3 bases sortantes. La ligation par la T4 DNA ligase peut se faire au travers d'un gap d'un nucléotide. L'analyse dans des conditions de déplétion de LigIV et XRCC4 (Kapusta et al., 2011) ou de DNA-PKcs (Malinsky et al, en préparation) montre que les cassures, bien que non réparées, sont maturées. En effet le dernier nucléotide en 5' est systématiquement éliminé lorsque la ligation finale est bloquée, même si celui-ci est compatible avec l'extrémité MAC sortante à l'autre borne.

Les résultats préliminaires obtenus lors de l'étude de la protéine CtIP semble indiquer qu'elle pourrait être responsable de la maturation des extrémités, au moins pour une partie des IES (figure 59). Cette hypothèse est cohérente avec le rôle de nucléase proposé pour CtIP. Chez *Paramecium tetraurelia*, la protéine homologue de CtIP est certainement impliquée dans la recombinaison méiotique puisqu'une déplétion de cette protéine entraîne des problèmes de méiose lors du passage de l'autogamie (voir résultats); elle pourrait avoir un rôle additionnel pour l'excision des IES lors de la maturation des extrémités d'ADN. Bien sur l'analyse que j'ai entreprise doit être étendue à d'autres IES car à l'heure actuelle, ce résultat n'a été

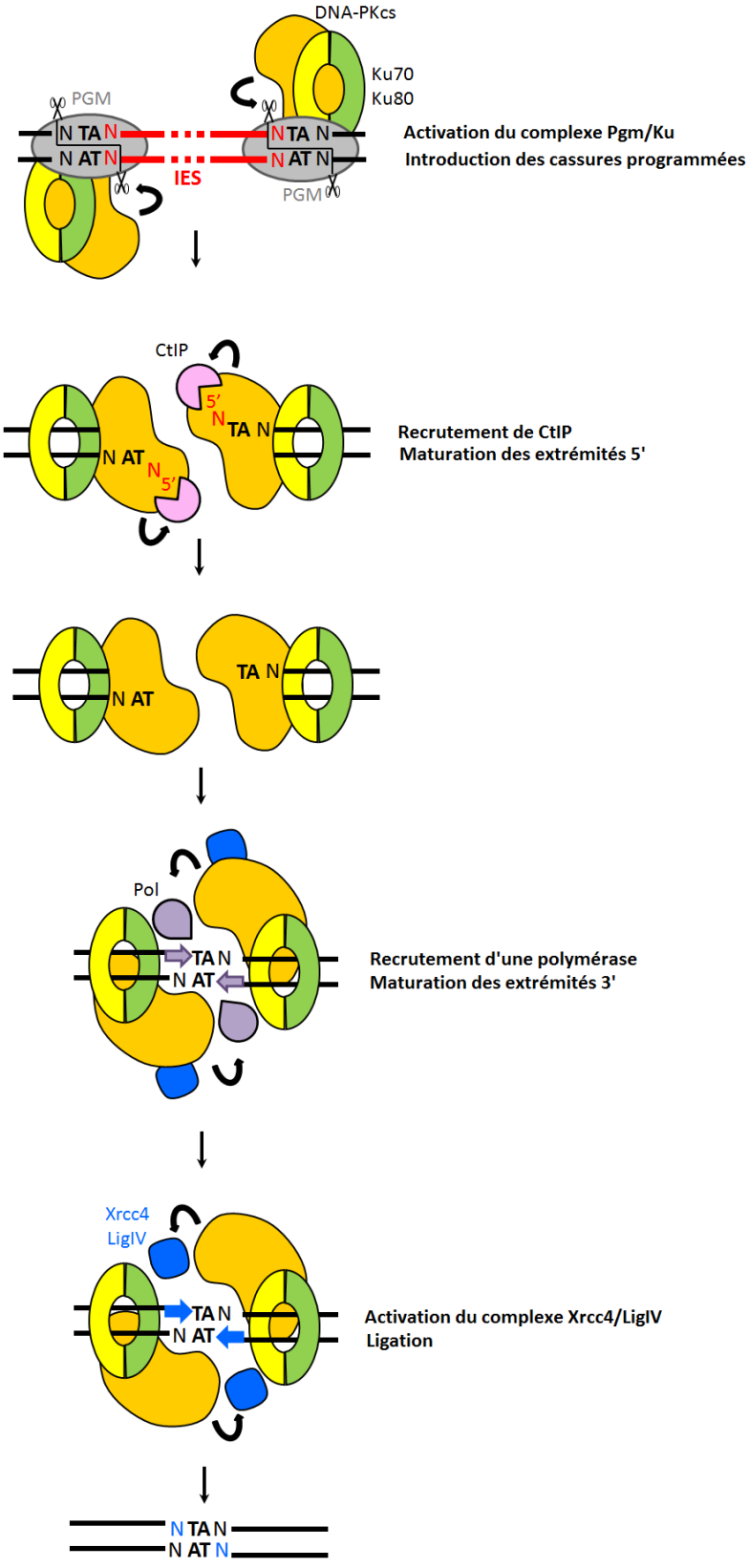


Figure 60 : Modèle pour le rôle de plate forme activatrice de DNA-PKcs, impliquée tout au long du processus d'excision des IES via son domaine kinase. Les flèches noires indiquent les phosphorylations putatives sur les acteurs de la voie d'excision des IES.

observé que pour une seule IES, sur deux testées. On pourrait aussi regarder au niveau des extrémités des IES. Il semble cependant que l'efficacité de formation des cercles excisés soit moindre dans des cellules déplétées pour CtIP par rapport au contrôle.

Enfin, les expériences d'extension à la terminale transférase permettent d'analyser la séquence au niveau des extrémités 3' et d'étudier la maturation de cette extrémité lors de l'excision des IES. Lorsque LigIV et/ou XRCC4 sont déplétées (Kapusta et al., 2011), j'ai pu montrer qu'on n'observait jamais l'ajout du nucléotide complémentaire en 3'. Cela indique que la maturation des extrémités d'ADN lors de l'excision des IES est partiellement dépendante du complexe de ligation. Cela suggère également un rôle précoce de la Ligase IV pendant le NHEJ avant même l'étape de ligation finale.

Il a été observé récemment (Cottarel et al., 2013) in vitro que l'autophosphorylation, de la DNA-PKcs, qui intervient normalement après la formation de la synapse, est dépendante de la présence de la LigaseIV, même si celle-ci n'a pas d'activité catalytique. La DNA-PKcs a de multiples rôles et de nombreuses cibles. L'un de ces rôles supposés est le recrutement des facteurs de maturation de l'ADN lors de la réparation par Non Homologous End Joining.

Le défaut de remplissage du gap 3' observé dans les cellules ou l'expression de la Ligase IV est inactivée pourrait s'expliquer de deux manières. Soit par un couplage entre la ligase IV et le recrutement (ou l'activation) de la polymérase qui mature les extrémités. Soit via l'action de LigIV sur la stimulation de l'activité kinase de la DNA-PKcs, qui aurait un rôle de plate forme activatrice lors de l'excision des IES (figure 60).

Chez la paramécie, la DNA-PKcs est également impliquée dans les réarrangements programmés du génome (Malinsky et al, en préparation). Dans les cellules déplétées pour cette protéine, l'excision est affectée pour une partie des IES, notamment les IES les plus longues. La déplétion en DNA-PKcs semble affecter toutefois l'introduction des cassures aux bornes des IES, on observe une accumulation de forme IES+ non excisées, et l'efficacité de réparation des cassures, on observe une persistance de CDB non réparées et une diminution de la quantité de jonctions d'excision. L'inactivation de DNA-PKcs n'a pas d'effet sur l'addition du nucléotide en 3'. Peut-être y a-t-il un effet quantitatif que les techniques employées ne nous permettent pas de voir : une inactivation partielle de DNA-PKcs pourrait seulement réduire la maturation des extrémités 3', permettant la formation de quelques jonctions d'excision.

Un processus intégré pour l'excision des IES

Une nouvelle vision de la voie NHEJ classique

Les résultats récents dans le domaine et l'analyse de la ligase IV chez *Paramecium tetraurelia* change progressivement la vision que l'on a de la voie NHEJ classique, qui semble ne semble pas fonctionner aussi simplement que par recrutement successif des différents acteurs du NHEJ (schématiquement CDB => KU => DNAPKcs => maturation => LigIV/XRCC4) :

La présence de l'hétérodimère Ku contrôle étroitement la maturation des extrémités. En plus de son rôle protecteur des extrémités, l'association de Ku et DNA-PKcs induit le changement de conformation qui va activer la protéine DNA-PKcs, qui peut elle-même recruter les facteurs de maturation de la voie NHEJ.

La maturation des extrémités est une étape cruciale pour le choix de la voie de réparation. Le niveau de résection des extrémités ADN contrôle en partie le choix de la voie de réparation.

Il a été montré in vitro que la présence de la ligase IV, mais pas son activité catalytique, est nécessaire pour l'autophosphorylation du complexe DNA-PK (Cottarel et al., 2013). Chez la paramécie, la ligase IV est nécessaire lors de l'excision des IES de la paramécie pour la maturation des extrémités d'ADN. Peut-être la protéine DNA-PKcs recrute un facteur de maturation des extrémités pour enlever le dernier nucléotide en 5'.

Des résultats récents ont montré que les protéines XRCC4 et XLF formaient des filaments. Cette structure particulière pourrait leur permettre d'aider à la formation de la synapse au niveau de la cassure double brin et diriger le recrutement de protéines de la voie NHEJ (Andres et al., 2012; Hammel et al., 2011; Ropars et al., 2011).

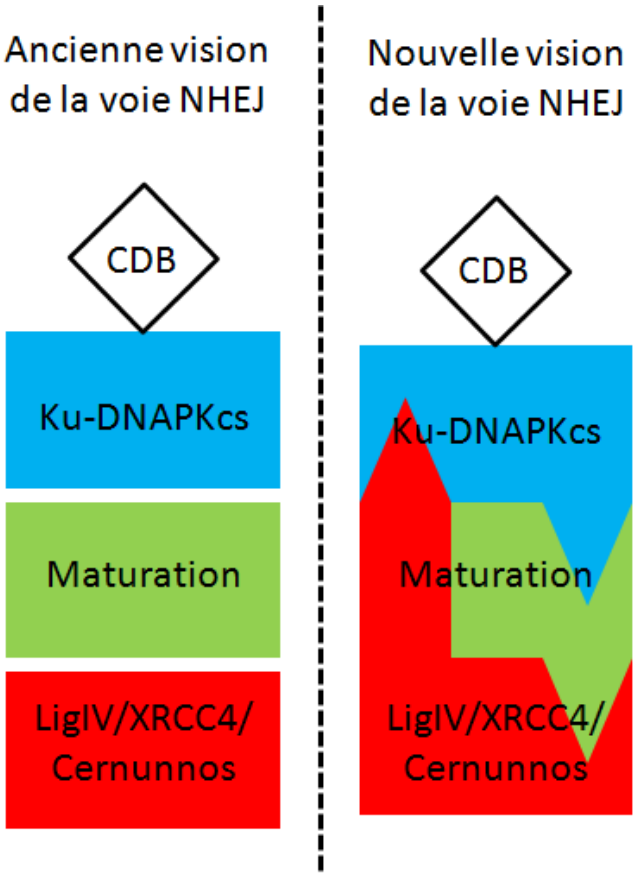


Figure 61 : Les deux visions de la voie NHEJ. Le recrutement séquentiel des différents acteurs à gauche. Le processus intégré où les différentes étapes sont étroitement liées à droite.

Il semble ainsi que la voie NHEJ soit un processus intégré, les acteurs en aval de la voie pouvant être requis pour des étapes en amont (figure 61). La protéine DNA-PKcs occupe un rôle central dans ce modèle, son activité kinase, elle-même dépendante de la présence d'autres acteurs de la voie NHEJ, peut permettre le recrutement d'autres acteurs, comme des nucléases ou des polymérases spécialisées, permettant une grande adaptabilité de la voie NHEJ.

L'excision des IES est un processus hautement intégré.

Si l'on parcourt la littérature, il y a une longue histoire de relation entre l'hétérodimère Ku et des éléments mobiles, particulièrement les transposons et les transposases :

Ku est connu pour interagir avec les éléments mobiles, il peut localiser au niveau des extrémités d'ADN viraux, et il a été observé, avant même la découverte du rôle de Ku au niveau des pieds du transposon P de drosophile (BEALL et al., 1994; Beall and Rio, 1996).

Ku interagit avec la transposase de *SleepingBeauty* et est nécessaire pour la transposition de cet élément mobile sans qu'on le sache toutefois si Ku est nécessaire pour l'excision de l'élément ou la réparation après intégration (Izsvak et al., 2004).

Lors de la recombinaison V(D)J, des indices suggèrent une interaction entre Ku et la transposase domestiquée RAG1. Cela avait amené les auteurs de l'étude à proposer un lien entre les deux étapes de la recombinaison V(D)J, introduction des cassures et réparation. En positionnant l'hétérodimère Ku à proximité des séquences RSS avant qu'elles soient clivées, les protéines RAG permettraient que Ku soit prêt à se fixer aux cassures double brin introduites et assurer une réparation par la voie NHEJ. Cependant l'absence de Ku n'empêche

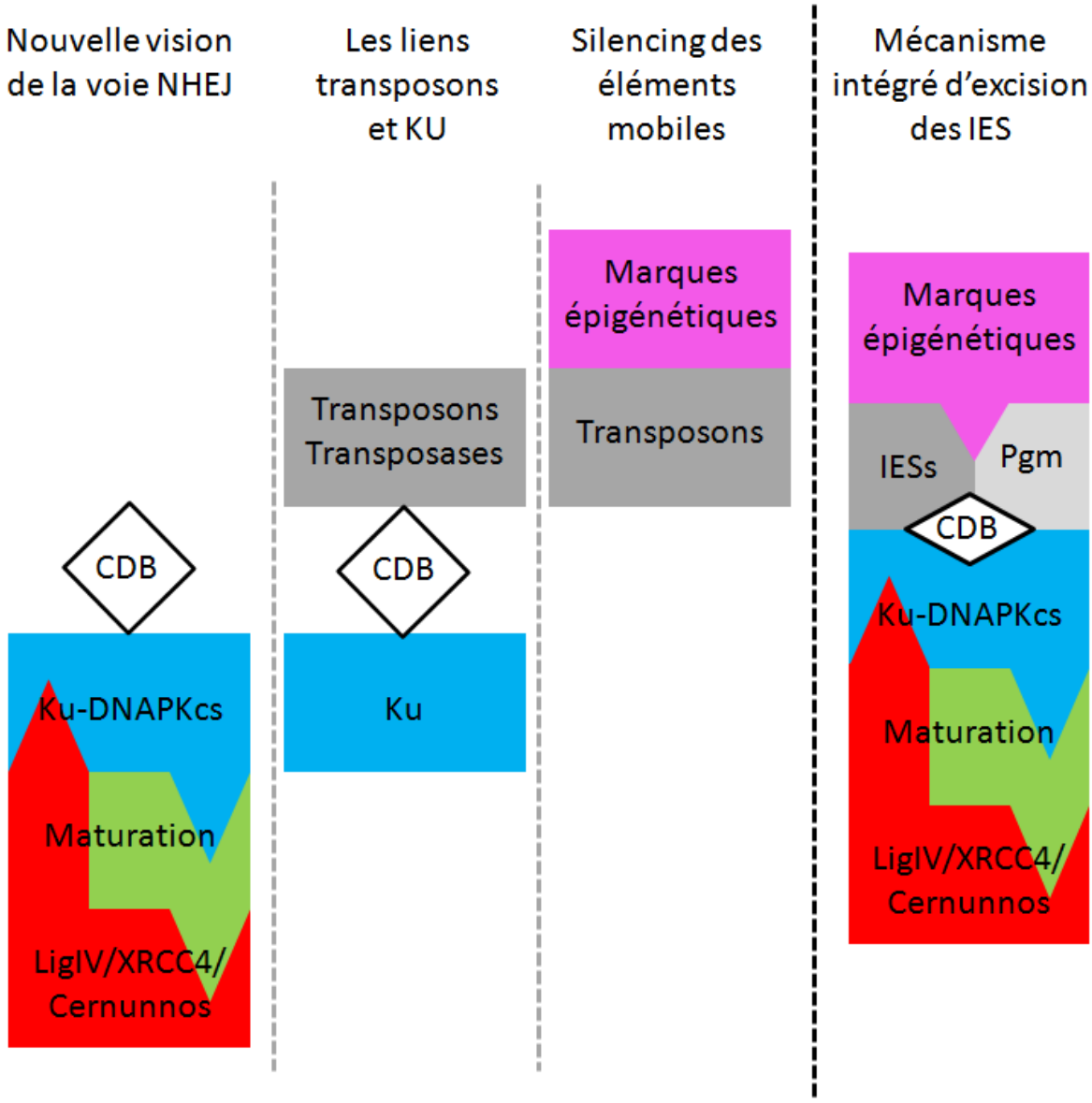


Figure 62 : Un processus hautement intégré pour l'excision des IES. Du marquage des séquences issues de transposons Tc1/mariner, éliminée par une transposase domestiquée de type piggybac, jusqu'à la réparation des cassures double brin programmées par la voie NHEJ.

pas l'introduction des cassures lors de la recombinaison V(D)J. L'existence d'usines de réarrangements nucléées autour des protéines RAG a même été proposée (Matthews and Oettinger, 2009). Le même genre de système est décrit chez *Tetrahymena* où en absence de Ku les cassures sont introduites normalement mais les *foci* où localisent les protéines impliquées dans les réarrangements sont désorganisés.

Chez *Paramecium tetraurelia*, la cellule a sélectionné au cours de l'évolution un mécanisme hautement intégré, qui regroupe « Ciblage des séquences à éliminer », « Introduction des cassures par une transposase domestiquée » et « recrutement de la voie NHEJ », le lien entre ces deux dernières étapes étant assuré par la protéine Ku (figure 62).

Comparaison avec Tetrahymena thermophila

Chez le cilié *Tetrahymena thermophila*, il y a 2 gènes *KU70*, *TKU70-1* et *TKU70-2* et un gène *KU80*, *TKU80*. De façon surprenante, en plus du gène de la transposase domestiquée Tpb2, il y a dans le génome un autre gène Tpb1, apparentée à une transposase *PiggyBac*, qui contient un domaine β-barrel de Ku qui sert de berceau à l'ADN et médie l'association entre les deux sous unités Ku70 et Ku80 (figure 63) (Cheng et al., 2010; Lin et al., 2012)

Chez *Tetrahymena*, l'excision des IES implique également la voie NHEJ. L'analyse des gènes Ku a été menée pendant les réarrangements programmés du génome. Si TKU80 est effectivement nécessaire pendant les réarrangements, les résultats sont différents de ceux obtenus chez *Paramecium tetraurelia.*

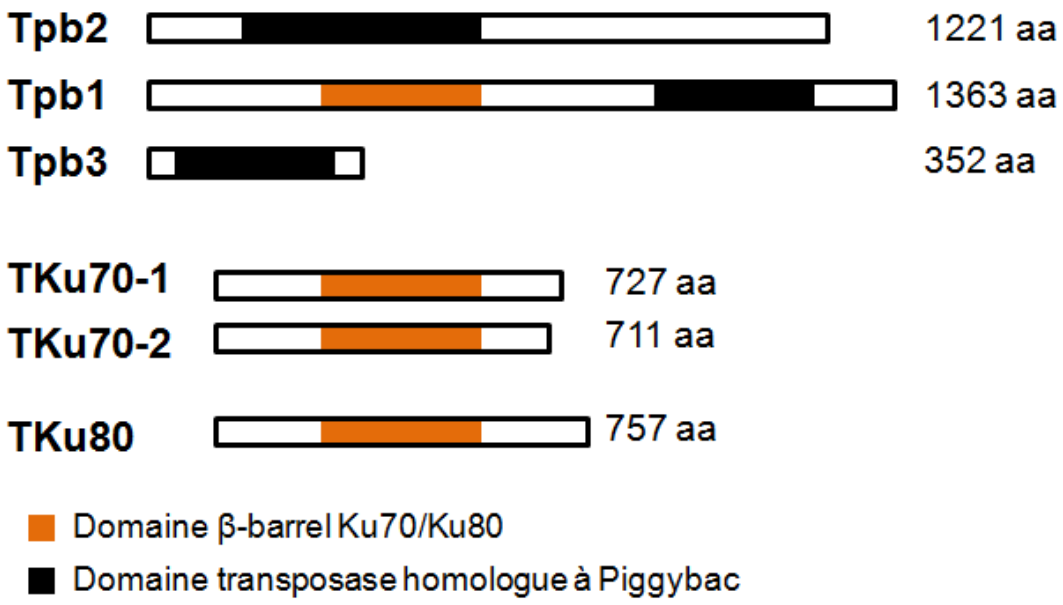


Figure 63 : Les différentes transposases domestiquées et protéines Ku chez *Tetrahymena thermophila*. Les domaines homologues à la transposase de Piggybac sont indiqués en noir. Les domaines β-barrel de Ku sont indiqués en orange.

Dans un KO *TKU80*, les cassures sont introduites normalement par Tpb2 et sont détectées par test TUNEL (Lin et al., 2012). Les auteurs ont été incapables de détecter les extrémités ADN par des techniques de biologie moléculaire tel que la LMPCR. Peut-être sont-elles dégradées ? On ne voit pas de jonctions d'excision MAC *de novo* mais des formes circulaires excisées des IES sont toujours détectées, ce qui appuie l'hypothèse selon laquelle les cassures sont introduites normalement. Il n'y a pas d'addition de télomères. Cette dernière observation est intéressante car elle n'indique pas forcément qu'il s'agit d'un défaut de coupure mais peut-être est-ce l'indice d'un rôle de Ku dans l'addition des télomères, comme cela peut-être observé chez la levure (Pfingsten et al., 2012).

Si l'absence de TKu80 n'empêche pas l'introduction des cassures, elle perturbe la formation des *foci* de Pdd1 dans le macronoyau en développement. C'est dans ces *foci* que localisent la grande majorité des protéines impliquées dans les réarrangements, dont Tpb2. Ces résultats ont mené les auteurs à proposer un modèle proche de celui proposé pour la recombinaison V(D)J : un lien qui tiendrait à proximité TKu80 et Tpb2, qui seraient impliqués dans de vastes « usines » de réarrangements.

Pourquoi un mécanisme différent chez *Tetrahymena thermophila ?*

De nombreux parallèles peuvent être faits lors des réarrangements programmés du génome chez *Paramecium* et *Tetrahymena.* Les deux ciliés doivent éliminer des séquences spécifiques de la lignée germinale. Tout deux recrutent une transposase domestiquée de type *piggybac* pour introduire les cassures double brin programmées. Tout deux utilisent la voie NHEJ pour réparer ces cassures double brin. Cependant la paramécie a sélectionné au cours de l'évolution un mécanisme rendant l'introduction des cassures par Pgm entièrement dépendante de la

réparation par la voie NHEJ. La première étape ne pouvant se faire sans un acteur majeur de la seconde, l'hétérodimère Ku. Pourquoi cette différence ? Peut-être est-ce dû aux différences qui existent entre les éléments éliminés chez l'un et l'autre cilié.

Chez *Paramecium*, plusieurs dizaines de milliers d'IES doivent être éliminées de façon extrêmement précise car elles interrompent des cadres ouverts de lecture. Chez *Tetrahymena*, les IES sont moins nombreuses, environ 6000, et n'interrompent pas les cadres ouverts de lecture. On observe aussi que la jonction de réparation macronucléaire n'est pas homogène. Il semble que cela soit du à une variabilité dans l'introduction des cassures double brin aux bornes des IES (Saveliev and Cox, 2001).

On peut imaginer que cette différence ait été le moteur de la divergence entre *Paramecium* et *Tetrahymena*. Soumis à de fortes pression de sélection sur la précision et l'efficacité de l'élimination des IES, *Paramecium* pourrait avoir fait un pas supplémentaire par rapport à *Tetrahymena* en couplant introduction des cassures et réparation, faisant de Ku un prérequis absolument indispensable pour introduire les cassures double brin programmées.

L'identification d'IES interrompant des cadres ouverts de lecture chez *Tetrahymena* est presque passée inaperçue lors de la publication de l'analyse des sites d'excision d'IES sur environ 25% du génome (Fass et al., 2011). En effet cela ne concerne qu'une dizaine de séquences pour cette portion du génome analysée. Mais trois de ces séquences sont bornées par des répétitions TTAA et semblent éliminées précisément. Par rapport aux autres séquences, elles sont presque anecdotiques mais on peut rapprocher cette observation d'une autre. *Tetrahymena* possède plusieurs gènes de type Piggybac dans son génome. L'analyse extensive des propriétés de Tpb2 est en cours dans plusieurs laboratoires (Cheng et al., 2010).

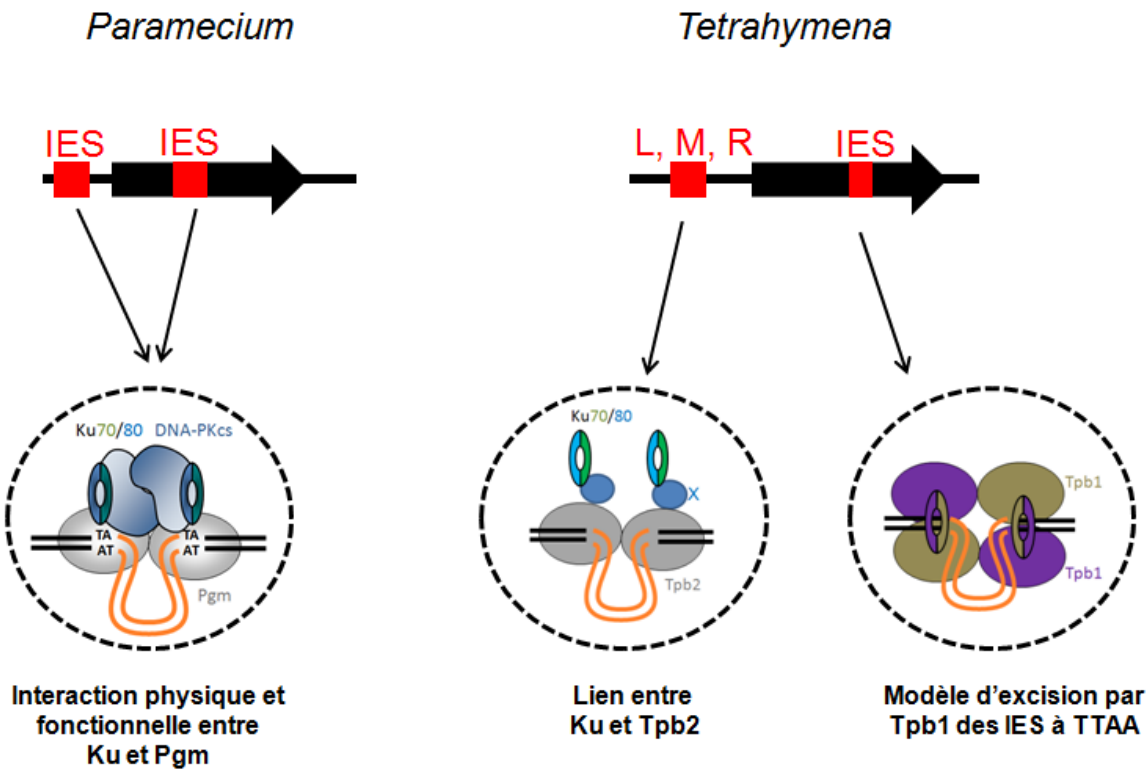


Figure 64 : Comparaison des mécanismes d'excision des IES chez *Paramecium tetraurelia* et *Tetrahymena thermophila*, et modèle d'excision des IES interrompant des cadres ouverts de lecture via Tpb1. Les éléments à éliminer sont indiqués en rouge. Les cadres ouverts de lecture par les flèches noires.

Mais un autre gène homologue, Tpb1, a été identifié. Ce gène, de façon très surprenante possède un domaine β-barrel de protéine Ku. J'ai observé un très bon alignement entre les domaines β-barrel des Ku80 de *Paramecium* et *Tetrahymena* et ce domaine de Tpb1. Il serait très intéressant de regarder si l'excision de ces IES particulières est dépendante de Ku et si Tpb1 a quelquechose à voir avec leur excision (figure 64). De plus, il existe d'autres protéines Tpb2-like chez *Tetrahymena*(Yao et al., 2007).

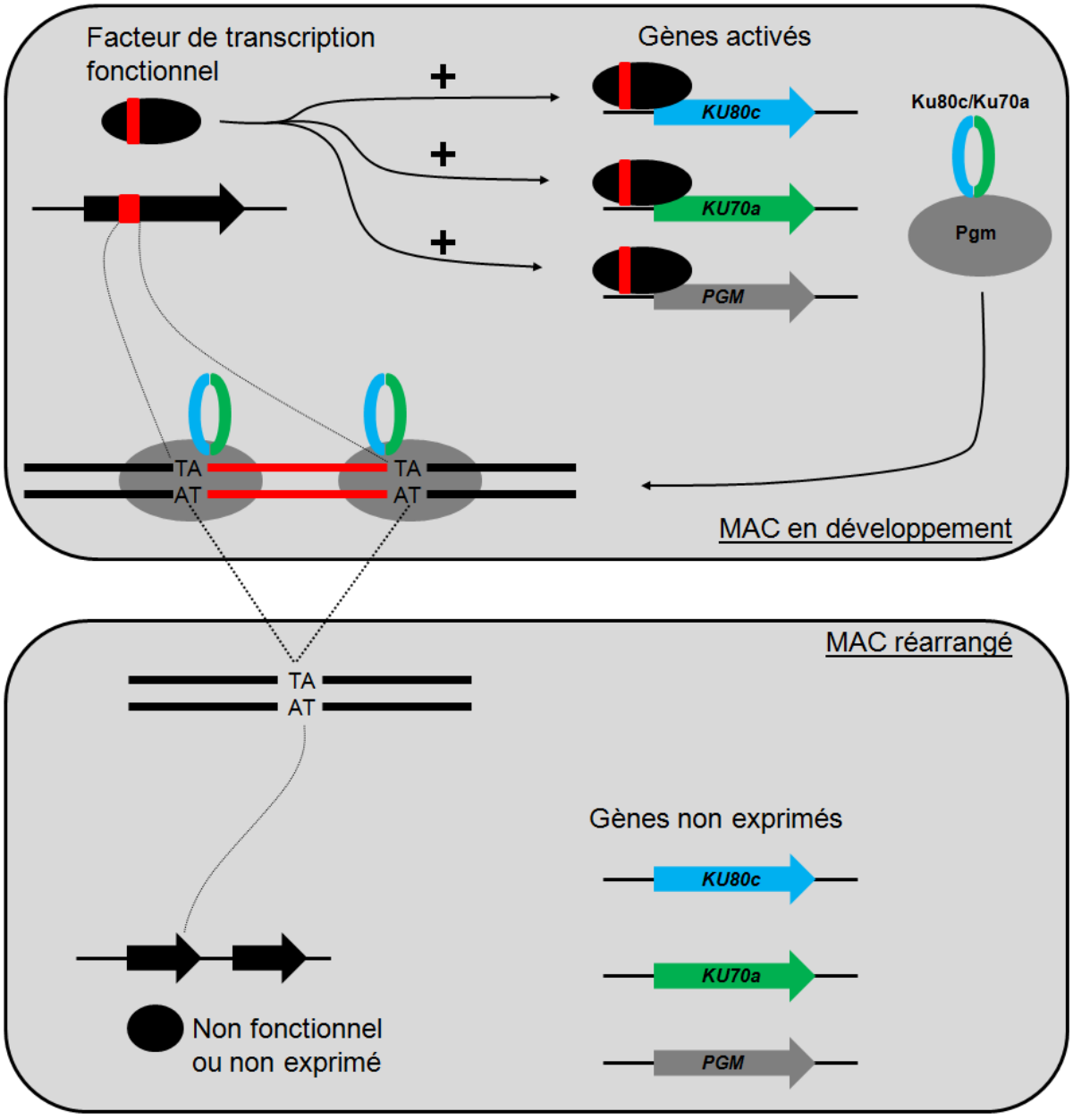


Figure 65 : Modèle de boucle de rétrocontrôle de l'expression des gènes impliqués dans les réarrangements programmés du génome.

Domestication du mécanisme d'excision des IES

Boucle de contrôle du programme des réarrangements

L'expression de Pgm et des acteurs de la voie NHEJ est programmé pendant les réarrangements du génome. En effet les protéines du complexe de ligation, *LIGIV* et *XRCC4* ont un pic de transcription pendant la méiose (Kapusta et al., 2011). *KU80c*, *KU70a* et *PGM* sont eux induits plus tardivement pendant le développement du macronoyau et les réarrangements du génome proprement dit. DNA-PKcs est induite au moment de la méiose, puis la transcription se poursuit pendant le développement du MAC (Malinsky et al, en préparation). De plus on a vu que le gène *KU80c* n'était pas transcrit en réponse à l'introduction des cassures par Pgm. En effet dans des cellules déplétées de Pgm, le gène *KU80c* est toujours induit. De plus le gène est toujours transcrit même à des temps tardifs, alors que son expression diminue fortement à ce stade dans les contrôles. Réciproquement, lorsque *KU80c* est déplété, le gène *PGM* est induit normalement, mais demeure transcrit jusqu'à de temps tardifs également. Et cela est vu pour d'autres gènes, normalement induits sous forme de pic, impliqués dans les réarrangements (Communication personnelle).

Puisqu'on a montré que les produits des gènes Pgm et Ku80c étaient impliqués dans l'introduction de cassures programmées aux bornes des IES lors des réarrangements programmés du génome, cela suggère que le programme de transcription des gènes impliqués dans les réarrangements est contrôlé par les réarrangements eux-mêmes, dans une boucle de rétrocontrôle (figure 65).

On pourrait imaginer un répresseur de transcription exprimé à partir du MAC en développement qui nécessite l'excision d'une IES pour être exprimé.

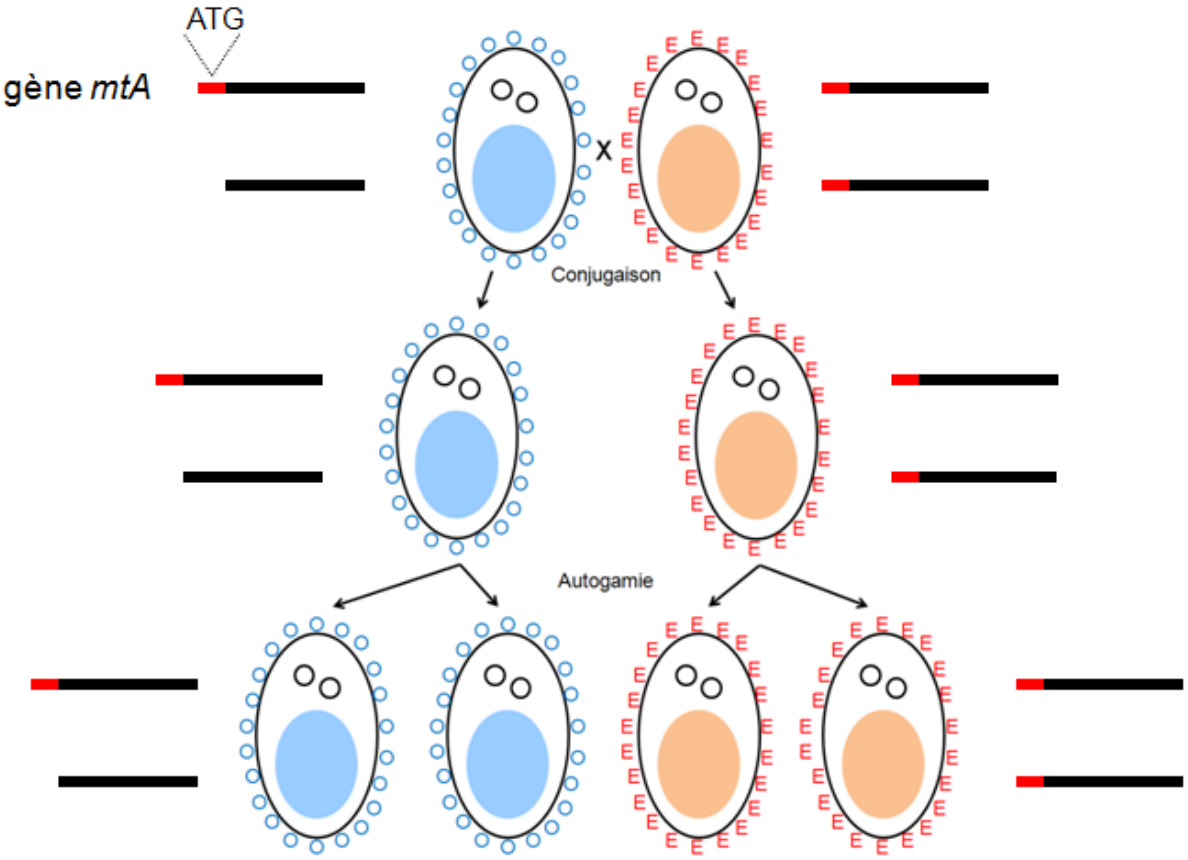


Figure 66 : Hérédité non mendelienne du mating type chez *Paramecium tetraurelia*. A coté est indiqué le gène *mtA* dans le mic ou le MAC des cellules. En rouge l'IES contenant le codon d'initiation du gène *mtA*.

On peut aussi imaginer un facteur de transcription spécifique des gènes des réarrangements programmés du génome, exprimé depuis les macronoyaux en développement qui ne serait actif ou fonctionnel que tant qu'une IES est présente. Ainsi les réarrangements programmés pourraient fournir un modèle élégant où l'élimination des séquences spécifiques de la lignée germinale éteindrait l'expression de ce facteur de transcription putatif, mettant fin naturellement à l'induction des gènes impliqués dans les réarrangements.

Chez les ciliés, il existe des précédents :

Chez *Oxytricha trifallax,* les gènes de transposase, qui sont retrouvés dans la lignée germinale ont un rôle clé lors des réarrangements programmés du génome. L'élimination des milliers de transposons de la lignée germinale est assurée par les transposases portées par ces transposons, qui sont induites pendant les réarrangements. Lorsque ces transposons sont éliminés, la transposase n'est plus exprimé et le système d'élimination des transposons s'éteint naturellement. C'est le mécanisme de mutualisme (Nowacki et al., 2009).

Chez *Euplotes*, il existe trois gènes de télomérase. L'un d'entre eux, *TERT-2*, est absent du génome macronucléaire et est spécifiquement induit pendant les réarrangements programmés du génome. Ce gène est exprimé à partir du nouveau MAC en développement avant d'être éliminé (Karamysheva et al., 2003).

La détermination du type sexuel, O ou E, chez *Paramecium tetraurelia* a longtemps été un mystère. En effet, le type sexuel est déterminé maternellement (figure 66). Les cellules issues d'un parent O seront toujours de type sexuel O, tandis que les cellules issues d'un parent E seront de type sexuel E. Depuis peu on sait que cette détermination est dépendante de la

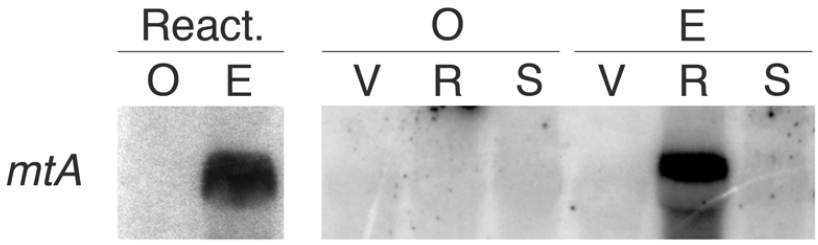


Figure 67 : Transcription du gène *mtA* dans des cellules de type sexuel O et E. V, cellules végétatives. R, cellules réactives. S, cellules en carence alimentaire. Les ARNs de ces cellules ont été extraits et une sonde marquée radioactivement révèle spécifiquement le gène *mtA*

présence ou pas d'une IES dans le gène *mtA.* En effet il a été observé que ce gène est spécifiquement transcrit dans les cellules dites « réactives », c'est à dire les cellules prêtes pour la conjugaison (figure 67). On trouve une IES au début de ce gène, cette IES contient le codon d'initiation du gène *mtA.* Sans cette IES le gène *mtA* n'est pas fonctionnel.
Les cellules de type sexuel E maintiennent cet IES dans leur génome macronucléaire tandis que les cellules de type sexuel O l'éliminent. L'élimination des IES étant contrôlée de façon épigénétique, l'élimination où la rétention de l'IES dépend du parent (Singh *et al*, en préparation).

J'ai présenté plus haut les travaux menés par le laboratoire de Sandra Duharcourt sur les centromères de *Paramecium tetraurelia.* Dans cette situation, je n'imagine pas que les IES ou les transposons, éliminés par Pgm, soient devenus des centromères spécifiques de la lignée germinale. J'imagine plutôt que les centromères sont des séquences ADN présentes avant la domestication de Pgm qui ont fini par être reconnue par le mécanisme d'élimination des séquences spécifiques de la lignée germinale.

Un candidat pour le contrôle des réarrangements programmés

Pour trouver un gène candidat qui pourrait contrôler l'expression des réarrangements programmés du génome, l'équipe de Linda Sperling a recherché par approche bioinformatique les IES qui étaient les plus transcrites dans les données de séquençage des ARNs messager pendant les premières phases du développement des macronoyaux. Parmi les meilleurs résultats on trouve un candidat prometteur. La rétention d'une IES pourrait modifier le cadre ouvert de lecture, décalant le codon stop d'environ 300 nucléotides en aval.

La réannotation du cadre ouvert de lecture fusionné prédit l'existence d'un domaine de liaison à l'ADN de type Myb-like qui pourrait être impliqué dans la terminaison de la transcription ou le remodelage de la chromatine (Cyril Denby Wilkes, communication personnelle).

Vérifier l'existence d'un messager pleine taille correspondant et tester le rôle de ce gène par ARN interférence pendant les réarrangements permettrait de confirmer cette prédiction bioinformatique et peut-être fournir un nouvel exemple de cooptation d'une séquence ADN macronucléaire par la machinerie d'excision des IES permettant le développement de nouvelles fonctions.

Le NHEJ alternatif pour l'élimination hétérogène des séquences répétées ?

Si je reviens à la première question posée au début de ma thèse, à savoir l'élimination hétérogène des séquences répétées, telles que des transposons ou des minisatellites, implique-t-elle l'utilisation d'une voie de recollement des extrémités indépendante de Ku ? Ku étant nécessaire pour l'introduction des cassures programmées par Pgm, je n'ai pas la réponse.

En effet les résultats obtenus montrent que Ku est un prérequis absolu pour l'introduction de cassures aux bornes des IES, il est donc impossible d'analyser le rôle de Ku lors de la réparation au niveau des jonctions macronucléaires puisque il n'y a pas d'extrémités ADN à analyser.

En ce qui concerne l'élimination hétérogène, Ku semble également impliqué dans le processus. Lorsque l'expression de Ku est éteinte, j'ai observé que les réarrangements imprécis sont inhibés. La forme longue du chromosome en aval du gène G, non réarrangée, est accumulée. Ce qui est cohérent avec un rôle couplé de Ku et Pgm pendant les réarrangements. En effet la transposase domestiquée est elle aussi indispensable pour les deux types de réarrangements (Baudry et al., 2009).

J'ai choisi au cours de ma thèse de m'orienter surtout vers l'analyse des cassures aux bornes des IES, je n'ai donc que peu de résultats à montrer sur le rôle éventuel de Ku dans les choix entre télomérisation des extrémités ou end joining lors de la fragmentation des chromosomes.

Avant de tester l'implication de Ku il est nécessaire de caractériser une région de réarrangements imprécis alternatifs du type de celle qui avait été décrite chez *Paramecium*

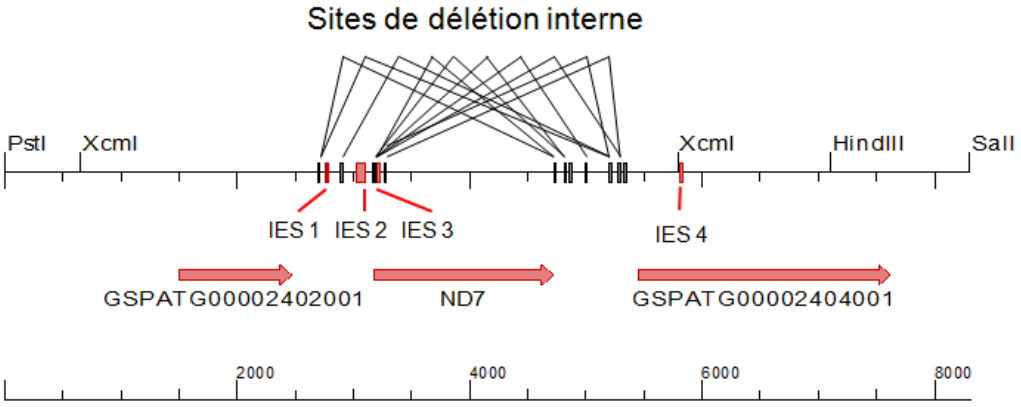


Figure 68 : Sites de délétions internes du gène *ND7* déterminés par PCR dans (Garnier et al., 2004)

primaurelia (Le Mouel et al., 2003). Chez *Paramecium tetraurelia*, comme il n'existe pas de régions naturelles de réarrangements imprécis caractérisée, je me suis concentré sur une région de délétion imprécise induite et héritable, celle d'un gène non essentiel *ND7*, déjà étudiée auparavant et qui semble se comporter tout à fait comme une région de réarrangements imprécis. Les jonctions macronucléaires au niveau de délétion induite du gène *ND7* montre une hétérogénéité au niveau de la réparation, une forte densité en TA, peut-être du fait de l'usage de microhomologie pour réparer la cassure. En ce qui concerne la fragmentation du chromosome à cet endroit, rien n'était connu (Garnier et al., 2004). J'ai commencé à caractériser cette région par des approches complémentaires d'électrophorèse en champ pulsé et en conditions classiques.

J'ai comparé l'ADN de cellules délétée ou non du gène *ND7*. J'ai réalisé le Southern Blot d'un fragment SalI/PstI avec le gène *ND7* en son centre permet de visualiser la délétion interne du gène. La méthode d'électrophorèse en champ pulsé, qui ne permet pas une résolution suffisante pour visualiser une délétion interne est en revanche adaptée à l'analyse de la fragmentation du chromosome portant le gène *ND7*. Les premiers résultats apportés par ces deux techniques montrent que la région de délétion imprécise induite du gène *nd7* se comporte comme une région de délétion imprécise naturelle, c'est-à-dire qu'on observe une délétion intrachromosomique et des formes fragmentées des chromosomes avec addition de télomères aux extrémités.

Une autre approche pour l'étude des réarrangements imprécis serait d'analyser les produits de réparation en utilisant la méthode de PCR. Un couple d'oligonucléotides placés de part et d'autre de la délétion interne du gène *ND7* a déjà été utilisé pour amplifier des produits de réparation (figure 68) (Garnier et al., 2004). Cette méthode montre des produits discrets

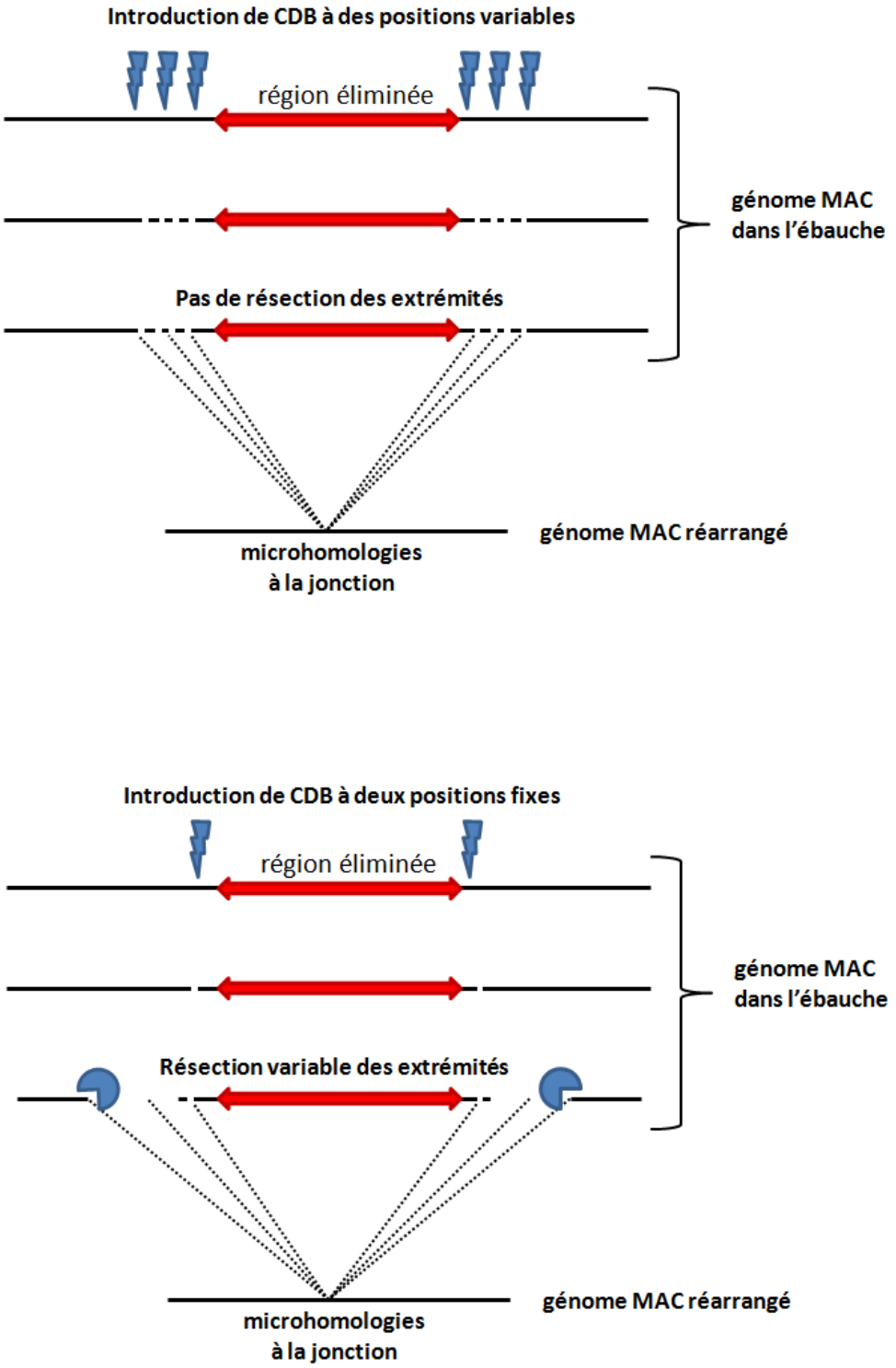


Figure 69 : Modèles pour les réarrangements imprécis pouvant expliquer les microhomologies observées au niveau des jonctions macronucléaires. En haut les cassures double brin sont introduites à des positions variables et réparées précisément par la voie NHEJ classique. En bas les cassures sont introduites à deux positions fixes et réparées imprécisément par la voie NHEJ alternative.

lorsqu'on les fait migrer sur gel, suggérant des formes préférentielles de réparation. L'analyse des jonctions de réparation montre qu'elles se situent dans les régions intergéniques en 5' et 3' de *ND7*, avec une tendance à se faire à proximité des bornes des IES. Ainsi les IES pourraient constituer des points chauds d'introduction de cassures pour les réarrangements imprécis.

Le séquençage des produits obtenus grâce aux PCRs télomériques permettrait de regarder si le point d'addition des télomères a également une tendance à être à proximité des bornes d'IES.

Etant donné que Ku est nécessaire également pour les réarrangements imprécis, son absence menant à l'amplification d'ADN non réarrangé, on pourrait imaginer un modèle ou l'imprécision de la délétion peut être due à l'introduction des CDB à des emplacements variables, couplée à une réparation sans dégradation des extrémités cassées Ku dépendante, le NHEJ (figure 69).

On ne peut pas exclure un modèle plus compliqué ou les cassures sont introduites par le tandem Pgm-Ku à des positions fixes, couplée à une réparation impliquant des étapes de résection des extrémités et Ku indépendante, le alt-NHEJ. Mais cette hypothèse semble peu probable, j'ai du mal à imaginer que la protéine Ku présente pour introduire les cassures double brin ne puisse pas assurer son rôle de protection des extrémités dans ce cas.

Bien sur ces deux hypothèses ne sont pas exclusives et on peut imaginer d'autres modèles plus compliqués.

L'analyse du rôle de Ku dans la réparation par NHEJ et de ce qui se passerait pendant l'excision des IES si l'hétérodimère était absent est bloquée par le rôle double de Ku, dans l'introduction des cassures et leur réparation. Cependant le développement de nucléases de

nouvelle génération, permettant de décider la séquence qui va être ciblée pourrait permettre de lever ce problème (Christian et al., 2010; Mahfouz et al., 2010). S'il était possible d'introduire grâce à ces nucléases une cassure au niveau d'une borne d'IES en condition d'extinction pour Ku, on pourrait enfin tester l'implication de la réparation alt-NHEJ dans les réarrangements.

Annexes

DETECTION DES CASSURES DOUBLE BRIN PAR LMPCR

1/ Préparation des linkers (20 pmol/µl)

dans un tube PCR de 0,2 ml:

oligo I' 100pmol/µl (5'-gctcggaccgtggctagcattagtc-3')	8µl
oligo (xTAy)J' 100pmol/µl (5'-xTAygactaatgcta-3')	8µl
Tris 1M pH 7,4	10µl
H2O	14µl

cycles de PCR: 5 minutes 95°C
10 secondes 70°C
redescendre à 4°C (à raison d'1°C par minute)

Le linker peut être conservé à -20°C plusieurs mois.

Dans les étapes suivantes, NE JAMAIS VORTEXER LES ECHANTILLONS

2/ Ligation du linker

dans un tube de 1,5ml:

ADN génomique de Paramécie	280ng
10X T4 DNA ligase buffer (Promega)	2 µl
T4 DNA ligase (3U Promega)	1µL
linker 20 pmol/µL	0,3 µl
Eau	qsp 20µl

Incuber à 18°C overnight
Incuber 10 minutes à 65°C pour inactiver la ligase

3/ Purification sur colonne Quiagen (voir ***protocole 2***)

Récupération de 50µl d'éluat (solution élution EB)
Les échantillons peuvent être conservés à -20°C

4/ PCR

dans un tube PCR de 0,2ml

éluat	10µl
DyNAzyme 10X buffer	2,5µl
oligo PCRhaut 10µM (5'-gaattcggatccgctcggaccgtggc-3')	1µl
oligo spécifique 10µM	1µl
dNTP 5mM chacun	1µl
H2O	9µl
DyNAzyme (1U)	0,5µl

cycles PCR:	2 minutes	95°C	
	1 minute	95°C	⎫
	1 minute	Tm	⎬ 35 cycles
	1 minute	72°C	⎭
	3 minutes	72°C	
	12°C		

5/ vérification sur gel

réalisation d'un gel d'agarose Nusieve 3% en TBE1X
dépôt de 5µl d'échantillon + 1µl de bleu
migration en TBE 0,5X

6/ Précipitation éthanol

reprendre les culots lyophilisés dans 15µl de 1X Taq Sequencing grade buffer (2 mM Mg2+ Promega)
Les échantillons peuvent être conservés à -20°C

7/ Préparation d'amorces marquées (à partir d'1µl de primer on réalise 10µl d'amorce*)

Dans un tube PCR de 0,2ml

primer 10µM	1µl
H2O	5 µl

incubation 2 minutes à 95°C
passage sur glace
Ajouter

γ^{33}P ATP (2500 Ci/mmole)	2,5µl
T4 PNK 10X buffer	1µl
kinase (10U/µl)	0,5µl

8/ Extension d'amorce

dans un tube PCR de 0,2 ml

10X sequencing grade buffer	1µl
dNTP 5mMe	1µl
amorce marquée	0,5µl
H2O	7,3µl
Taq sequencing grade (5U/µL)	0,2µl
echantillons	15µl

cycles PCR:

2 minutes	95°C	
1 minute	95°C	|
1 minute	Tm	| 1-10 cycles (fonction des résultats du gel BET)
1 minute	72°C	|
3 minutes	72°C	
12°C		

9/ Gel de séquence

dépôt de 3,5 µl de chaque échantillon

DETECTION D'EXTREMITES 3'OH LIBRES PAR TdT

EXTENSION A LA TERMINAL TRANSFERASE :

1°) Dans un tube PCR de 0,2 ml mélanger :

500 ng	ADN génomique (*dosé sur gel par coloration Bet)*
qsp **154,8 µl**	H_2O

2°) Dénaturation 3 min à 95°C (*pour dégager les extrémités 3'0H*). Transfert dans la glace

3°) Ajouter :

20 µl	tampon NEB 4 10X
20 µl	$CoCl_2$ 2,5 mM (NEB)
1 µl	H_2O
1 µl	Terminal Transferase (20U NEB)
3,2 µl	dCTP ou dGTP 1 mM
45,2 µl	

3°) Incubation : 45 min à 37°C
10 min à 70°C
∞ à 4°C *(dans machine PCR).*

4°) Transfert dans un tube 1,5 ml. Précipitation EtOH avec 1/2 volume de NH_4OAc 7,5M et 1 µl glycogène 35 mg/ml.
Rinçage culot avec EtOH 70% et séchage 3 min au lyophilisateur

6°) Reprendre dans :

4 µl	tampon DynaZyme 10X (Finnzyme)
3,2 µl	oligo A(G) ou A(C) 10 µM (*en fonction de l'extension*)
2 µl	dNTP 5 mM each
30,8 µl	H_2O
40 µl	

AMPLIFICATION DU PRODUIT D'EXTENSION :

1°) Transfert dans tube 0,2 ml :
3 min à 95°C
30 min à 50°C
∞ à 4°

2°) ajouter :

1 µl	tp Dynazyme 10X (Finnzyme)
2 µl	oligo I 10 µM
2 µl	oligo 2 10 µM
4,5 µl	H_2O
0,5 µl	Dynazyme (1U Finnzyme)
10 µl	

3°) amplification PCR :

1 min à 50°C
3 min à 72°C

35 X	1 min	95°C
	1 min	60°C [1]
	3 min	72°C

7 min à 72°C
∞ à 4°C

VISUALISATION PAR EXTENSION D'AMORCE :

1°) Précipitation EtOH avec 1/2 volume de NH_4OAc 7,5M (***sans rajouter de glycogène supplémentaire)***
Rinçage culot avec EtOH 70% et séchage 3 min au lyophilisateur.

2°) Reprendre le culot dans 20 µl tp Sequencing grade 1X (50mM Tris pH9, 2 mM Mg^{2+})

3°) A **10 µl** d'échantillon, ajouter :

1,5 µl	tp Sequencing grade 10X
1 µl	dNTP 5 mMe
0,5 µl	amorce marquée au ^{33}P (cf annexe plus bas, kit fmol)
11,8 µl	H_2O
0,2 µl	Taq Sequencing Grade (5U/µl Promega)
15 µl	

4°) extension :

hot start 2 min à 95°C

1 à 10 X [2]	1 min à 95°C
	1 min à xx°C (*en fonction du Tm de l'amorce)*
	7 min à 72°C

∞ à 4°C

[1] à ajuster en fonction du Tm des oligos 2 et I
[2] à ajuster en fonction de l'efficacité de la réaction. Un cycle suffit en général.

EXTRACTION D'ADN DE PARAMECIE ET INSERTS

A partir de 400 ml de culture de variété 4 en WGP1X + ß à environ 1000 cellules par ml (ou 200 mL de culture autogame à 2000-5000 ç/mL).

1°) Filtrer la culture à travers de la gaze pliée en 8 plis puis centrifuger 4 x 100 ml en poires (100 mL/poire), à 1100 rpm pendant 1 min à RT (*Bien sécher l'extérieur des poires, sinon elles restent coincées dans les godets. Vérifier que les godets sont bien placés dans la centrifugeuse. Décompter le temps à partir du moment où la vitesse est atteinte)*.

2°) Mettre immédiatement des pipettes bouchées dans le fond des poires (*pour empêcher les cellules de remonter en suspension*) et éliminer le surnageant.

3°) Resuspendre les cellules dans le restant de milieu et tout rassembler dans une seule poire. Compléter avec du Tris 10 mM pH 7,4 pour rinçage.

4°) Centrifuger comme en 1°)

5°) Bloquer à nouveau les cellules avec une pipette bouchée, repérer le volume du culot et éliminer le surnageant.

6°) Centrifuger à nouveau comme en 1°)

7°) Eliminer le surnageant à la pipette Pasteur en laissant le culot dans 1,2 mL de volume final (*graduations de la poire*). Resuspendre le culot et prélever une goutte pour lame de microscopie

1 - INSERTS

8°) transférer 300 µL de cellules (*vol à ajuster en fonction de la quantité de cellules*) dans 2 mL d'agarose low melting 1,5% maintenu en surfusion à 45-50°C dans un bain-marie (InCert agarose Cambrex #50121 dans EDTA 125 mM pH 9).

9°) Homogénéiser le mélange cellules + agarose à la pipette Pasteur et couler dans un moule à plugs. Transfert à 4°C le temps de finir l'extraction des ADN.

10°) transférer les plugs dans 7 mL de solution de lyse + protéinase K 1 mg/mL préchauffée à 50-55°C (Falcon 15). Incuber à 50-55°C pendant 15-24 heures (agiter manuellement au début).

11°) lavages des plugs 2 x 1h à 50°C dans 25 mL EDTA 0,1M pH9 (agiter manuellement de temps en temps).

12°) Stockage à 4°C dans le dernier rinçage.

2 - ADN GENOMIQUE

10°) Recentrifuger le restant des cellules dans la poire comme en 1°)

11°) éliminer le surnageant en laissant 1 volume de Tris au dessus du culot.

12°) Resuspendre le culot de cellules à la pipette Pasteur.

13°) Ajouter dans la poire 4 vol de culot de solution de lyse + protéinase K 1 mg/mL préchauffée à 50-55°C

14°) Transfert dans Falcon 15. Incuber à 50-55°C pendant 15-24 heures puis passage à 4°C.

15°) Ajouter 1 ml de phénol saturé avec du Tris 0,1 M pH 8 dans chaque tube. Agitation douce pendant 1 heure.

16°) Dialyses: 2 h dans 2 litres de TE 1X pH 8 EtOH 25%
Une nuit dans 2 litres de TE 1X pH8 EtOH 25%
2 h dans 2 litres de TE 1X pH 8

17°) Récupérer l'ADN dans un tube Eppendorf. Stockage à 4°C (*A chaque utilisation, il faudra mélanger le contenu du tube par inversions successives, puis le centrifuger pendant 1 minute*).

18°) Dosage à 260 nm d'une dilution 1:200^{e} de la prep (*attention au facteur 2 si on utilise une cuve de 0,5 cm de large*).

14°) Pour analyser l'ADN génomique, dépôt sur gel de l'équivalent de 10 µl d'une prep à 0,1 de DO_{260} (1:200^{e})

EXTRACTION TRIZOL : ARN et Protéines

PROTOCOLE prélèvement des cellules → billes

1) Filtrer à travers 8 plis de gaze le volume nécessaire (dépend de la concentration en cellules)

2) Centrifuger 1' RT 1100 à 1600 rpm les paramécies (dans n poires) – bloquer la remontée des paramécies avec baguette et vider le surnageant

3) Rassembler toutes les cellules dans 1 poire

4) Centrifuger et vider comme en 2

5) Centrifuger encore, et retirer le maximum de surnageant

6) Remettre les cellules en suspension

7) AZOTE LIQUIDE : préparer bac avec béchers baignant dans l'azote liquide.
Pipeter toutes les cellules, et laisser tomber les gouttes dans les béchers (billes). Si plus de billes que de bécher, attendre que la précédente descende au fond du bécher.

8) Récupérer les billes (pinces) et les mettre dans un cryotube rempli de et baignant dans l'azote liquide → -80°C.

PROTOCOLES D' EXTRACTIONS

MATERIEL ET REACTIFS

- Réactif TRIZOL (Gibco BRL)

- ou H_2O ARN de chez Jean

- $CHCl_3$, isopropanol et EtOH spécial ARN. Préparer du EtOH 75% extemporanément avec H_2O ARN

- Billes de verre 425-600 µm (Unwashed. Sigma) *(Lavées 3 x 15 min dans HCl 1N dans une bouteille de 250 ml. Rinçages abondants avec H2O milliQ jusqu'à ce que le pH = celui de l'eau. Cuisson 2 hr à 200-250°C dans une étuve Memmert).*

- tubes Eppendorf et Falcon 15 propres, Pipetmans pour ARN, cônes à filtres

- Nettoyer la paillasse et l'extérieur de tous les flacons/boîtes/Pipetman avec un Kleenex imbibé de SDS 1%, sans rincer.

- Centrifugeuse Jouan réglée à 4°C

PROTOCOLE ARN : (*pour < 4 x 10^5 cellules. Marche encore pour 200 ml de culture à 3000-4000 cellules/ml*)

1) Sortir les billes de paramécie dans la carboglace

2) Dans un tube Falcon 50, peser 4g de billes de verre, puis transférer les billes dans un tube Falcon 15. Sous la hotte, ajouter 4mL de Trizol

3) Transférer les cellules dans le Trizol tout en vortexant le tube (basse vitesse) (*lyse des cellules*).

4) Incubation 5' à RT (*on peut faire une série de tubes et minuter à partir du dernier*)

5) Ajouter 800µL de $CHCl_3$ par tube. Vortex vitesse maxi 15 sec. Incubation 2-3' à RT.

6) Centrifugation 15' à 4°C (Jouan, vitesse max) → *deux phases + galette d'ADN entre*.

7) Récupérer le surnageant (~ 2,4mL) et répartir 3 x 800µL en tubes Eppendorf.
GARDER les tubes (contenant les billes) pour extraction des protéines.

8) Dans chaque tube, ajouter 670µL d'isopropanol (propanol 2). Homogénéiser et centrifuger 10' à 4°C et à 15K (*gros culots blancs*).

9) 2 rinçages avec 1mL EtOH 75%. Vortex si le culot ne se décolle pas, et centrifugation 5' à 7500g (9,8K) à 4°C. *1^{er} rinçage (→ vortex) → centri et $2^{ème}$ rinçage → vortex → -20°C conservation OK*.

10) Séchage des culots à l'air libre. Resuspension *(peut être très longue)* des ARN dans 30µL H_2O + ARN pour chaque tube (*10 min à 55-60°C peuvent aider*), et pooler les échantillons.

11) Stockage des ARN à –80°C

12) Mesure à 260 et 280 nm de la DO d'une dilution $1/250^{ème}$ dans du TE (ou H_2O). Pour les cuves Eppendorf jetables, 75 µl d'échantillon suffisent (1 U DO_{260} = 40 µg d'ARN)

13) Pour un Northern : 20 µg d'ARN total par piste.

PROTOCOLE PROTÉINES :

1) Recupérer la phase phénol et l'interphase de l'étape 7) (environ 2ml), répartir en trois tubes eppendorf. *(peut être gardé la nuit à 4°C)*

2) Ajouter 400µL d'éthanol absolu. Mélanger par inversion.

3) Incubation 2-3' à RT

4) Centrifugation à 2000g pendant 5’

5) Récupérer le surnageant et répartir en 6 tubes

6) Précipiter en ajoutant 1 mL d’isopropanol/tube. Incubation 10’ à RT

7) Centrifugation 10’ à 12000g à 4°C.

8) TROIS lavages dans 1mL de 0,3M guanidine hydrochloride dans 95 % éthanol. Chaque lavage 20’ à RT puis 5’ de centri à 7500g. *(On peut conserver les protéines pendant plusieurs mois à -20°C dans 0,3M guanidine hydrochloride dans 95 % éthanol)*

9) Après dernier lavage rincer à l’éthanol 20’ à RT puis centrifuger 5’ à 7500g.

10) Sécher le culot protéique au speed-vac.

11) Solubiliser les culots dans du SDS 1% en pipettant. La dissolution complète peut nécessiter une incubation à 50°C.

REFERENCES BIBLIOGRAPHIQUES

Ahnesorg, P., Smith, P., and Jackson, S.P. (2006). XLF interacts with the XRCC4-DNA ligase IV complex to promote DNA nonhomologous end-joining. Cell *124*, 301-313.

Andres, S.N., Vergnes, A., Ristic, D., Wyman, C., Modesti, M., and Junop, M. (2012). A human XRCC4–XLF complex bridges DNA. Nucleic acids research.

Arnaiz, O., Mathy, N., Baudry, C., Malinsky, S., Aury, J.-M., Wilkes, C.D., Garnier, O., Labadie, K., Lauderdale, B.E., Mouel, A.L.*, et al.* (2012). The Paramecium Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences. PLoS genetics.

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N.*, et al.* (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature *444*, 171-178.

Barnes, D.E., Stamp, G., Rosewell, I., Denzel, A., and Lindahl, T. (1998). Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. Curr Biol *8*, 1395-1398.

Bassing, C.H., Swat, W., and Alt, F.W. (2002). The Mechanism and Regulation of Chromosomal V(D)J Recombination. Cell.

Baudat, F., and de Massy, B. (2004). [SPO11: an activity that promotes DNA breaks required for meiosis]. Med Sci (Paris) *20*, 213-218.

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., and Betermier, M. (2009). PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes & development *23*, 2478-2483.

BEALL, E.L., ADMON, A., and RIO, D.C. (1994). A Drosophila protein homologous to the human p70 Ku autoimmune antigen interacts with the P transposable element inverted repeats. Proc Natl Acad Sci.

Beall, E.L., and Rio, D.C. (1996). Drosophila IRBP/Ku p70 corresponds to the mutagen-sensitive mus309 gene and is involved in P-element excision in vivo. Genes & development.

Berger, J.D. (1973). Nuclear differentiation and nucleic acid synthesis in well-fed exconjugants of Paramecium aurelia. Chromosoma *42*, 247-268.

Boboila, C., Yan, C., Wesemann, D.R., Jankovic, M., Wang, J.H., Manis, J., Nussenzweig, A., Nussenzweig, M., and Alt, F.W. (2010). Alternative end-joining catalyzes class switch recombination in the absence of both Ku70 and DNA ligase 4. The Journal of experimental medicine *207*, 417-427.

Bouhouche, K., Gout, J.-F., Kapusta, A., Betermier, M., and Meyer, E. (2011). Functional specialization of Piwi proteins in Paramecium tetraurelia from post-transcriptional gene silencing to genome remodelling. Nucleic acids research.

Bracht, J.R., Perlman, D.H., and Landweber, L.F. (2012). Cytosine methylation and hydroxymethylation mark DNA for elimination in Oxytricha trifallax. Genome biology.

Budman, J., Kim, S.A., and Chu, G. (2007). Processing of DNA for nonhomologous end-joining is controlled by kinase activity and XRCC4/ligase IV. The Journal of biological chemistry *282*, 11950-11959.

Chan, C.Y., Galli, A., and Schiestl, R.H. (2008). Pol3 is involved in nonhomologous end-joining in Saccharomyces cerevisiae. DNA repair *7*, 1531-1541.

Cheng, C.-Y., Vogt, A., Mochizuki, K., and Yao, M.-C. (2010). A Domesticated piggyBac Transposase Plays Key Roles in Heterochromatin Dynamics and DNA Cleavage during Programmed DNA Deletion in Tetrahymena thermophila. Molecular biology of the cell.

Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., and Voytas, D.F. (2010). Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. Genetics.

Ciubotaru, M., Trexler, A.J., Spiridon, L.N., Surleac, M.D., Rhoades, E., Petrescu, A.J., and Schatz, D.G. (2013). RAG and HMGB1 create a large bend in the 23RSS in the V(D)J recombination synaptic complexes. Nucleic acids research.

Corneo, B., Wendland, R.L., Deriano, L., Cui, X., Klein, I.A., Wong, S.Y., Arnal, S., Holub, A.J., Weller, G.R., Pancake, B.A*., et al.* (2007). Rag mutations reveal robust alternative end joining. Nature *449*, 483-486.

Corpet, A., and Almouzni, G. (2009). A histone code for the DNA damage response in mammalian cells? The EMBO journal *28*, 1828-1830.

Cottarel, J., Frit, P., Bombarde, O., Salles, B., Négrel, A., Bernard, S., Jeggo, P.A., Lieber, M.R., Modesti, M., and Calsou, P. (2013). A noncatalytic function of the ligation complex during nonhomologous end joining. The Journal of cell biology.

Coyne, R.S., Lhuillier-Akakpo, M., and Duharcourt, S. (2012). RNA-guided DNA rearrangements in ciliates: Is the best genome defence a good offence? Biol Cell.

Daley, J.M., Palmbos, P.L., Wu, D., and Wilson, T.E. (2005). Nonhomologous end joining in yeast. Annual review of genetics *39*, 431-451.

Deriano, L., Stracker, T.H., Baker, A., Petrini, J.H., and Roth, D.B. (2009). Roles for NBS1 in alternative nonhomologous end-joining of V(D)J recombination intermediates. Molecular cell *34*, 13-25.

Doak, T.G., Cavalcanti, A.R.O., Stover, N.A., Dunn, D.M., Weiss, R., Herrick, G., and Landweber, L.F. (2003). Sequencing the Oxytricha trifallax macronuclear genome: a pilot project. Trends Genet.

Dubois, E., Bischerour, J., Marmignon, A., Mathy, N., Regnier, V., and Betermier, M. (2012). Transposon Invasion of the Paramecium Germline Genome Countered by a Domesticated

PiggyBac Transposase and the NHEJ Pathway. International Journal of Evolutionary Biology.

Duharcourt, S., Butler, A., and Meyer, E. (1995). Epigenetic self-regulation of developmental excision of an internal eliminated sequence on Paramecium tetraurelia. Genes & development *9*, 2065-2077.

Duharcourt, S., Keller, A.M., and Meyer, E. (1998). Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in Paramecium tetraurelia. Molecular and cellular biology *18*, 7075-7085.

Duharcourt, S., Lepere, G., and Meyer, E. (2009). Developmental genome rearrangements in ciliates: a natural genomic subtraction mediated by non-coding transcripts. Trends Genet *25*, 344-350.

Fang, W., Wang, X., Bracht, J.R., Nowacki, M., and Landweber, L.F. (2012). Piwi-Interacting RNAs Protect DNA against Loss during Oxytricha Genome Rearrangement. Cell.

Fass, J.N., Joshi, N.A., Couvillion, M.T., Bowen, J., Hamilton, E.P., Orias, E., Hong, K., Coyne, R.S., Eisen, J.A., Chalker, D.L.*, et al.* (2011). Genome-Scale Analysis of Programmed DNA Elimination Sites in Tetrahymena thermophila. Genes Genomes Genetics.

Fisher, T.S., and Zakian, V.A. (2005). Ku: A multifunctional protein involved in telomere maintenance. DNA repair.

Garnier, O., Serrano, V., Duharcourt, S., and Meyer, E. (2004). RNA-mediated programming of developmental genome rearrangements in Paramecium tetraurelia. Molecular and cellular biology *24*, 7370-7379.

Goodarzi, A.A., Yu, Y., Riballo, E., Douglas, P., Walker, S.A., Ye, R., Harer, C., Marchetti, C., Morrice, N., Jeggo, P.A.*, et al.* (2006). DNA-PK autophosphorylation facilitates Artemis endonuclease activity. The EMBO journal *25*, 3880-3889.

Gratias, A., and Betermier, M. (2001). Developmentally programmed excision of internal DNA sequences in Paramecium aurelia. Biochimie *83*, 1009-1022.

Grawunder, U., Wilm, M., Wu, X., Kulesza, P., Wilson, T.E., Mann, M., and Lieber, M.R. (1997). Activity of DNA ligase IV stimulated by complex formation with XRCC4 protein in mammalian cells. Nature *388*, 492-495.

Guirouilh-Barbat, J., Huck, S., Bertrand, P., Pirzio, L., Desmaze, C., Sabatier, L., and Lopez, B.S. (2004). Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells. Molecular cell *14*, 611-623.

Hammel, M., Rey, M., Yu, Y., Mani, R.S., Classen, S., Liu, M., Pique, M.E., Fang, S., Mahaney, B.L., Weinfeld, M.*, et al.* (2011). XRCC4 Protein Interactions with XRCC4-like Factor (XLF) Create an Extended Grooved Scaffold for DNA Ligation and Double Strand Break Repair. Journal of Biological Chemistry.

Huertas, P. (2010). DNA resection in eukaryotes: deciding how to fix the break. Nature structural & molecular biology *17*, 11-16.

Huertas, P., and Jackson, S.P. (2009). Human CtIP mediates cell cycle control of DNA end resection and double strand break repair. The Journal of biological chemistry *284*, 9558-9565.

Iacovoni, J.S., Caron, P., Lassadi, I., Nicolas, E., Massip, L., Trouche, D., and Legube, G. (2010). High-resolution profiling of γH2AX around DNA double strand breaks in the mammalian genome. The EMBO journal.

Indiviglio, S.M., and Bertuch, A.A. (2009). Ku's essential role in keeping telomeres intact. Proceedings of the National Academy of Sciences of the United States of America *106*, 12217-12218.

Izsvak, Z., Stuwe, E.E., Fiedler, D., Katzer, A., Jeggo, P.A., and Ivics, Z. (2004). Healing the Wounds Inflicted by Sleeping Beauty Transposition by Double-Strand Break Repair in Mammalian Somatic Cells.

Jahn, C.L., and Klobutcher, L.A. (2002). Genome remodeling in ciliated protozoa. Annual review of microbiology *56*, 489-520.

Jaraczewski, and Jahn (1993). Elimination of Tec elements involves a novel excision process. Genes & development.

Kapusta, A., Matsuda, A., Marmignon, A., Ku, M., Silve, A., Meyer, E., Forney, J.D., Malinsky, S., and Bétermier, M. (2011). Highly Precise and Developmentally Programmed Genome Assembly in Paramecium Requires Ligase IV–Dependent End Joining. PLoS genetics.

Karamysheva, Z., Wang, L., Shrode, T., Bednenko, J., Hurley, L.A., and Shippen, D.E. (2003). Developmentally Programmed Gene Elimination in Euplotes crassus Facilitates a Switch in the Telomerase Catalytic Subunit. Cell.

Karrer, K.M., and VanNuland, T.A. (2002). Methylation of adenine in the nuclear DNA of Tetrahymena is internucleosomal and independent of histone H1. Nucleic acids research.

Klobutcher, Turner, and LaPlante (1993). Circular forms of developmentally excised DNA in Euplotes crassus have a heteroduplex junction. Genes & development.

Klobutcher, L.A., and Herrick, G. (1995). Consensus inverted terminal repeat sequence of Paramecium IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons. Nucleic acids research *23*, 2006-2013.

Klobutcher, L.A., and Herrick, G. (1997). Developmental genome reorganization in ciliated protozoa: the transposon link. Progress in nucleic acid research and molecular biology *56*, 1-62.

Le Mouel, A., Butler, A., Caron, F., and Meyer, E. (2003). Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in paramecia. Eukaryotic cell *2*, 1076-1090.

Lepere, G., Betermier, M., Meyer, E., and Duharcourt, S. (2008). Maternal noncoding transcripts antagonize the targeting of DNA elimination by scanRNAs in Paramecium tetraurelia. Genes & development *22*, 1501-1512.

Lepere, G., Nowacki, M., Serrano, V., Gout, J.F., Guglielmi, G., Duharcourt, S., and Meyer, E. (2009). Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia. Nucleic acids research *37*, 903-915.

Li, X., and Heyer, W.D. (2008). Homologous recombination in DNA repair and DNA damage tolerance. Cell research *18*, 99-113.

Lieber, M.R., Lu, H., Gu, J., and Schwarz, K. (2008). Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. Cell research *18*, 125-133.

Lin, I.-T., Chao, J.-L., and Yao, M.-C. (2012). An essential role for the DNA breakage-repair protein Ku80 in programmed DNA rearrangements in Tetrahymena thermophila. Molecular biology of the cell.

Lisby, M., Barlow, J.H., Burgess, R.C., and Rothstein, R. (2004). Choreography of the DNA damage response: spatiotemporal relationships among checkpoint and repair proteins. Cell *118*, 699-713.

Longhese, M.P., Bonetti, D., Guerini, I., Manfrini, N., and Clerici, M. (2009). DNA double-strand breaks in meiosis: checking their formation, processing and repair. DNA repair *8*, 1127-1138.

Lu, H., Shimazaki, N., Raval, P., Gu, J., Watanabe, G., Schwarz, K., Swanson, P.C., and Lieber, M.R. (2008). A biochemically defined system for coding joint formation in V(D)J recombination. Molecular cell *31*, 485-497.

Ma, J.L., Kim, E.M., Haber, J.E., and Lee, S.E. (2003). Yeast Mre11 and Rad1 proteins define a Ku-independent mechanism to repair double-strand breaks lacking overlapping end sequences. Molecular and cellular biology *23*, 8820-8828.

Ma, Y., Schwarz, K., and Lieber, M.R. (2005). The Artemis:DNA-PKcs endonuclease cleaves DNA loops, flaps, and gaps. DNA repair *4*, 845-851.

Mahfouz, M.M., Li, L., Shamimuzzaman, M., Wibowo, A., Fang, X., and Zhu, J.-K. (2010). De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. PNAS.

Matthews, A.G.W., and Oettinger, M.A. (2009). RAG: a recombinase diversified. Nature immunology.

Mimitou, E.P., and Symington, L.S. (2008). Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. Nature *455*, 770-774.

Mimitou, E.P., and Symington, L.S. (2009). DNA end resection: many nucleases make light work. DNA repair *8*, 983-995.

MIMORI, T., AKIZUKI, M., YAMAGATA, H., SHINICHI INADA, YOSHIDA, S., and HOMMA, M. (1981). Characterization of a High Molecular Weight Acidic Nuclear Protein Recognized by Autoantibodies in Sera from Patients with Polymyositis-Scleroderma Overlap. J Clin Invest.

Nowacki, M., Higgins, B.P., Maquilan, G.M., Swart, E.C., Doak, T.G., and Landweber, L.F. (2009). A functional role for transposases in a large eukaryotic genome. Science (New York, NY *324*, 935-938.

Nowacki, M., Zagorski-Ostoja, W., and Meyer, E. (2005). Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in Paramecium tetraurelia. Curr Biol *15*, 1616-1628.

Ochi, T., Sibanda, B.L., Wu, Q., Chirgadze, D.Y., Bolanos-Garcia, V.M., and Blundell, T.L. (2010). Structural Biology of DNA Repair: Spatial Organisation of the Multicomponent Complexes of Nonhomologous End Joining. Journal of Nucleic Acids.

Pfingsten, J.S., Goodrich, K.J., Taabazuing, C., Ouenzar, F., Chartrand, P., and Cech, T.R. (2012). Mutually Exclusive Binding of Telomerase RNA and DNA by Ku Alters Telomerase Recruitment Model. Cell.

Ropars, V., Drevet, P., Legrand, P., Baconnais, S., Amram, J., Faur, G., Márquez, J.A., Piétrement, O., Guerois, R., Callebaut, I.*, et al.* (2011). Structural characterization of filaments formed by human Xrcc4–Cernunnos/XLF complex involved in nonhomologous DNA end-joining. PNAS.

Rosidi, B., Wang, M., Wu, W., Sharma, A., Wang, H., and Iliakis, G. (2008). Histone H1 functions as a stimulatory factor in backup pathways of NHEJ. Nucleic acids research *36*, 1610-1623.

Rossetto, D., Truman, A.W., Kron, S.J., and Côté, J. (2010). Epigenetic Modifications in Double-Strand Break DNA Damage Signaling and Repair. Clin Cancer Res.

Saintigny, Y., Delacote, F., Boucher, D., Averbeck, D., and Lopez, B.S. (2007). XRCC4 in G1 suppresses homologous recombination in S/G2, in G1 checkpoint-defective cells. Oncogene *26*, 2769-2780.

Saveliev, and Cox (2001). Product analysis illuminates the final steps of IES deletion in Tetrahymena thermophila. The EMBO journal.

Shrivastav, M., De Haro, L.P., and Nickoloff, J.A. (2008). Regulation of DNA double-strand break repair pathway choice. Cell research *18*, 134-147.

Singleton, B.K., Torres-Arzayus, M.I., Rottinghaus, S.T., Taccioli, G.E., and Jeggo, P.A. (1999). The C terminus of Ku80 activates the DNA-dependent protein kinase catalytic subunit. Molecular and cellular biology *19*, 3267-3277.

Slijepcevic, P., and Al-Wahiby, S. (2005). Telomere biology: integrating chromosomal end protection with DNA damage response. Chromosoma *114*, 275-285.

Sontheimer (2012). Small RNAs of Opposite Sign but Same Absolute Value. Cell.

Strande, N., Roberts, S.A., Oh, S., Hendrickson, E.A., and Ramsden, D.A. (2012). Specificity of the dRP/AP Lyase of Ku Promotes Nonhomologous End Joining (NHEJ) Fidelity at Damaged Ends. The Journal of biological chemistry.

Swart, E.C., Bracht, J.R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J.S., Goldman, A.D., Nowacki, M., Schotanus, K.*, et al.* (2013). The Oxytricha trifallax Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes. Plos Biol.

Tomimatsu, N., Mukherjee, B., Deland, K., Kurimasa, A., Bolderson, E., Khanna, K.K., and Burma, S. (2012). Exo1 plays a major role in DNA end resection in humans and influences double-strand break repair and damage signaling decisions. DNA repair.

Walker, J.R., Corpina, R.A., and Goldberg, J. (2001). Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature *412*, 607-614.

Wang, H., Perrault, A.R., Takeda, Y., Qin, W., Wang, H., and Iliakis, G. (2003). Biochemical evidence for Ku-independent backup pathways of NHEJ. Nucleic acids research *31*, 5377-5388.

Wang, H., Rosidi, B., Perrault, R., Wang, M., Zhang, L., Windhofer, F., and Iliakis, G. (2005). DNA ligase III as a candidate component of backup pathways of nonhomologous end joining. Cancer research *65*, 4020-4030.

Wang, M., Wu, W., Wu, W., Rosidi, B., Zhang, L., Wang, H., and Iliakis, G. (2006). PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. Nucleic acids research *34*, 6170-6182.

Weterings, E., and Chen, D.J. (2008). The endless tale of non-homologous end-joining. Cell research *18*, 114-124.

Williams, K., G.Doak, T., and Herrick, G. (1993). Developmental precise excision of Oxytricha trifallax telomere-bearing elements and formation of circles closed by a copy of the flanking target duplication. The EMBO journal.

Wohlbold, L., and Fisher, R.P. (2009). Behind the wheel and under the hood: functions of cyclin-dependent kinases in response to DNA damage. DNA repair *8*, 1018-1024.

Yao, M.-C., Yao, C.-H., Halasz, L.M., Fuller, P., Rexer, C.H., Wang, S.H., Jain, R., Coyne, R.S., and Chalker, D.L. (2007). Identification of novel chromatin-associated proteins involved in programmed genome rearrangements in *Tetrahymena*. Journal of cell science.

Yu, X., and Gabriel, A. (2003). Ku-dependent and Ku-independent end-joining pathways lead to chromosomal rearrangements during double-strand break repair in Saccharomyces cerevisiae. Genetics *163*, 843-856.

Résumé.

L'élimination programmée d'ADN spécifique de la lignée germinale pour former un nouveau noyau somatique a été décrite chez les eucaryotes. Ces réarrangements sont initiés par l'introduction de cassures double brin (CDB) de l'ADN et la préservation de l'intégrité du génome requiert une réparation efficace. Chez *Paramecium tetraurelia*, le génome est largement réarrangé pendant le développement du nouveau noyau somatique, après l'introduction de milliers de cassures double brin programmées par la transposase domestiquée PiggyMac (Pgm)

Ces réarrangements consistent en l'excision précise de dizaines de milliers de séquences uniques et non codantes (IES) qui interrompent 47% des gènes dans la lignée germinale ; et l'élimination hétérogène de séquences répétées qui mène à des délétions internes de taille variable ou à la fragmentation des chromosomes avec addition de télomères aux extrémités.

L'implication de la voie du Non Homologous End Joining (NHEJ) dans l'excision précise des IES a été prouvée. Dans des cellules déplétées de Ligase IV ou XRCC4, les cassures aux bornes des IES sont introduites normalement mais il n'y a pas de jonctions d'excision formées et les extrémités cassées s'accumulent sans être dégradées. Mais la voie de réparation impliquée dans les réarrangements imprécis est encore inconnue. L'hypothèse d'une réparation par la voie NHEJ alternative (alt-NHEJ), indépendante de Ku et impliquant la résection des extrémités et l'utilisation de microhomologie, a été émise. C'est pourquoi pendant ma thèse je me suis intéressé à ma thèse au rôle des protéines Ku.

Deux gènes *KU70* et trois gènes *KU80* ont été identifiés dans le génome de la paramécie. *KU70a* et *KU80c* sont spécifiquement induits pendant les réarrangements programmés du génome et les protéines localisent dans les noyaux somatiques en développement. Des expériences d'extinction de ces gènes par ARN interférence ont prouvé que ces gènes étaient indispensables. Au niveau moléculaire, l'ADN non réarrangé est amplifié dans les cellules déplétées de Ku. De plus, les cassures double brin programmées ne sont pas introduites aux bornes des IES.

Mes résultats suggèrent que Ku fait partie d'un complexe de pré-excision, avec la transposase domestiquée Pgm, et est nécessaire pour l'introduction des cassures double brin programmées pendant les réarrangements programmés du génome.